ISSN: 3023 -8951

CHAOS AND FRACTALS

VOLUME 3, ISSUE 2, JULY 2026

AN INTERDISCIPLINARY JOURNAL OF NONLINEAR SCIENCE

ADBA

https://journals.adbascientific.com/chf

ISSN: 3023-8951

Chaos and Fractals

Volume: 3 – Issue No: 2 (July 2026)

EDITORIAL BOARD

https://journals.adbascientific.com/chf/about/editorialTeam

ISSN: 3023-8951

Chaos and Fractals

Volume: 3 – Issue No: 2 (July 2026)

CONTENTS

https://journals.adbascientific.com/chf/issue/view/22

ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.65-75
https://doi.org/10.69882/adba.chf.2026071

Descriptions of Cantor Sets: A Set-Theoretic Survey and Open Problems

Mohsen Soltanifar *,1
*Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, 620-155 College Street, Toronto, ON M5T 3M7, Canada.

ABSTRACT This paper synthesizes the principal descriptive set-theoretic perspectives on deterministic Cantor sets on the real line and charts directions for future study. After recounting their historical genesis and compiling an up-to-date taxonomy, we review the Borel hierarchy and four hierarchically ordered representations, general, nested, iterated-function-system (IFS), and $q-$ary expansion, presented from the most general to the most specific set-theoretic description of deterministic Cantor sets. We then present explicit closed-form descriptions for two thin families of measure-zero Cantor sets and, for the augmented "thick" family of positive Lebesgue measure, a new closed recursive formula for the endpoint maps governing every retained interval at every finite stage; we further show that the classical middle-third set lies in the intersection of all three families, providing a unifying example across the four representations.The survey closes by isolating several open problems in four directions, aiming to provide mathematicians with a coherent platform for further descriptive set-theoretic investigations into Cantor-type sets on the real line.

KEYWORDS
Cantor set
Borel operations
Formulas
Set theory
Fractal geometry

INTRODUCTION

The conceptual roots of the Cantor Sets lie in Georg Cantor's 1872 introduction of the *derived set* (the set of all limit points of a given subset of $\mathbb{R}$), which laid the groundwork for perfect, nowhere-dense subsets of $\mathbb{R}$. Shortly thereafter, H. J. S. Smith's 1875 note provided the first printed construction of a "Cantor-type" set by iteratively excising subintervals of $[0,1]$, thereby demonstrating that a closed set can be uncountable yet possess arbitrarily small outer content, although the significance of his example went largely unnoticed at the time. Cantor returned to the theme in his six–part memoir *Über unendliche lineare Punktmannichfaltigkeiten* (1879–1884); Part V (written October 1882, published 1883) contains the now-classical *3-ary or ternary expansion set* (later called the middle-third Cantor set), presented as a perfect set that is nowhere dense.

In a letter dated November 1883 Cantor extends this construction of the Cantor set by defining the *Cantor function*, a continuous, non-decreasing map that is constant on the removed intervals and thus furnishes a counter-example to naïve forms of the Fundamental Theorem of Calculus. His 1884 paper *De la puissance des ensembles parfaits de points* in *Acta Mathematica* analyzed the cardinalities of perfect sets, cementing the role of the ternary Cantor set as both a set-theoretic and analytic paradigm. These milestones, Smith's overlooked precursor, Cantor's derived-set framework, the explicit ternary construction, and the analytic Cantor function, together mark the emergence of what is now simply known as the Cantor set (Fleron 1994). Various classes and approaches to the construction of Cantor sets have been extensively studied, each generating distinct subclasses and mathematical structures. As of the date of this survey (31 January 2026), mathematicians have identified numerous specific types of Cantor sets, including Affine, Autonomous, Balanced, Bunched, Central, Cookie-Cutter, Deterministic, Discrete, Distorted, Dusts, Fat, Generalized, Genus-g, Geometric, Hairy, High Dimensional, Homogeneous, Hyperbolic, Iteration Function System (IFS) type, Invariant, Kinetic, Linear, Minimal, Multi-model, Nearly-affine, Non autonomous, Non-homogeneous, Non-archimedean (p-adic), Non-central, Non-Symmetric, Nonstandard, p-Adic Heisenberg, Random, Regular, Scrawny, Self-Similar, Semi-bounded, Smooth, Sticky, Superior, Smith-Volterra-Cantor(SVC) type, Symbolic, Symmetric, Tame, Thick, Thin, Twofold, Ultrametric, Uniform, Universal, Visible, and Wild. Comprehensive references for these subclasses can be found in (Edgar 2007; Vallin 2013; Falconer 2014; Gorodetski and Northrup 2015; Lin *et al.* 2024; McLoughlin and Krizan 2014; Jumha *et al.* 2024; Nassiri and Bidaki 2025). In this paper, however, our primary focus is on deterministic Cantor sets defined explicitly on the real line, and we direct the interested reader to relevant literature for more general or abstract variants.

Manuscript received: 11 March 2026,
Revised: 03 May 2026,
Accepted: 14 May 2026.

[1]mohsen.soltanifar@alumni.utoronto.ca (**Corresponding author**)

Since Georg Cantor's original construction in the 1880s, the Cantor set has become a touchstone throughout mathematics and beyond. On the real line it provides one of the simplest explicit exemplars of a fractal: it is nowhere dense (its closure has empty interior), totally disconnected (its only connected subsets are singletons), perfect (closed with no isolated points), and uncountable, yet its Lebesgue measure can be tuned from zero (the classical middle-third set) to any prescribed value in $(0, 1)$. Such extremal combinations of topological and measure-theoretic properties make Cantor sets a natural testing ground for "pathological" phenomena, most famously the Cantor–Lebesgue function, a continuous, non-decreasing map that is constant on a set of full measure.

In constructive and descriptive settings the Cantor set reappears as a paradigmatic building block: for example, basic Cantor-type function blocks span $5/28$ of the canonical representatives in the function space $F(R, R)$, (Soltanifar 2023) and they appear to prove the existence of aleph-two deterministic as well as aleph-two random fractals on the real line (Soltanifar 2021, 2022). Outside pure mathematics, Cantor-type patterns have been cited in art, early Egyptian column motifs (Frame *et al.* 2025; Kainth 2023), in modeling financial markets (Jin 2021), and in nature, such as ring-like stratification of Saturn when idealized as repeated circular products (Mandelbrot 1982).

Beyond the illustrations just listed, Cantor-type sets recur across mathematics as both structural prototypes and technical tools, and a brief tour of these connections clarifies why a unified descriptive survey is timely. The principal connections between them can be summarized as follows:

- Hyperbolic invariant sets of smooth dynamical systems frequently carry the topology of a Cantor set, and the conjugacy between such systems and full shifts on finite alphabets is the foundational technique of symbolic dynamics (Lind and Marcus 1995). The horseshoe construction of Smale and the wild hyperbolic sets of Newhouse (1979) are paradigmatic examples, and the IFS attractors discussed in Section are their simplest one-dimensional analogues.
- The set of badly approximable real numbers in $[0, 1]$ is, up to measure zero, a countable union of Cantor-type sets indexed by partial-quotient bounds. The arithmetic theory of sums and differences of Cantor sets, initiated by Newhouse and developed extensively by (Palis and Takens 1993; Moreira and Yoccoz 2001), asks when $\Gamma + \Gamma'$ or $\Gamma - \Gamma'$ contains an interval, with applications to homoclinic bifurcations and to the Markov and Lagrange spectra (Moreira *et al.* 2020).
- Self-similar measures supported on Cantor sets serve as test cases and counterexamples in Fourier analysis. Salem sets, compact sets whose Hausdorff and Fourier dimensions coincide, are constructed using randomized Cantor-type schemes (Salem 1951), and Cantor measures appear centrally in modern work on arithmetic progressions in sparse sets (Łaba and Pramanik 2009).
- Cantor sets are the canonical examples in the theory of Hausdorff dimension, self-similar measures, and ectifiability (Mattila 1995; Falconer 2014). The Hausdorff-Dimension Theorem (Soltanifar 2006b) is itself most transparently proved by exhibiting an explicit one-parameter family of Cantor sets, and the recent extensions in (Soltanifar 2021, 2022) place such constructions in a broader cardinality-theoretic framework.
- Brouwer's classical theorem characterizes the Cantor space $\{0, 1\}^{\mathbb{N}}$ as the unique (up to homeomorphism) non-empty perfect, compact, totally disconnected, metrizable space (Brouwer 1910; Kechris 1995). Every Cantor set on the real line is therefore homeomorphic to every other, which makes the descriptive distinctions surveyed in this paper, arising not from topology but from Lebesgue measure, Hausdorff dimension, and explicit arithmetization, the more striking.
- Random analogues of the constructions in Section give rise to random recursive fractals (Mauldin and Williams 1986; Soltanifar 2022) and to determinantal point processes whose intensity measures are concentrated on generalized Cantor sets (Lin *et al.* 2024). These constructions are the stochastic counterpart of the deterministic survey pursued here.
- In addition to the art-, finance-, and physics-related references cited above, Cantor-type structures have been invoked as descriptive models in signal processing and information theory (Schroeder 2009).

Several of these directions, particularly the arithmetic of intersections in dynamical systems and the harmonic-analytic study of Cantor measures, motivate the open problem (OP5) on common intersections.

Despite this wide relevance, no unified survey has yet cataloged the diverse set-theoretic descriptions of Cantor sets, nor identified which directions are fully developed and which remain open. The present paper closes that gap by reviewing known explicit and implicit set-theoretic formulaes on the real line for the deterministic cases, classifying them by structural features, and highlighting avenues for further investigation. While the book-length treatment of Vallin (2013) catalogs many properties and applications of Cantor sets, it does not organize their descriptions into a hierarchical set-theoretic framework, nor does it supply explicit endpoint recursions for positive-measure middle-α families; the present work is, to our knowledge, the first survey to do so.

Beyond consolidating known material, this survey advances the descriptive set-theoretic study of Cantor sets in four concrete directions viewed here:

- We organize the general, nested, iterated-function-system, and q–ary descriptions under a single set-theoretic language, making translation between representations routine.
- Theorem 0.3 supplies a closed recursive formula for the endpoint maps $a_{n,k}(p/q)$, $b_{n,k}(p/q)$ of the augmented thick family $\Gamma_3(p/q, 2)$ with $0 < p/q \leq 1/3$; to the best of our knowledge this is the first such explicit recursion in the literature.
- Corollary 0.4 identifies the classical Cantor set as a member of all three families simultaneously, with equivalence across all four representations made explicit.
- The accompanying `R` implementation (Appendix B) realizes the recursion in (N2) for exact finite-stage computation, and The present framework leads naturally to five open problems (OP1-OP5), each of which is amenable to computational experimentation.

These contributions distinguish the present work from prior catalogs and position it as a platform for further descriptive set-theoretic research on Cantor-type sets of the real line.

This paper focuses exclusively on deterministic Cantor sets on the real line, examined through a set-theoretic lens, and highlights several descriptive perspectives that remain undiscovered. The remainder of this paper is organized as follows. We first review the necessary background in descriptive set theory, recalling the Borel hierarchy and introducing four nested levels of set-theoretic representation, ranging from the broadest class of Cantor sets to the narrowest class of construction. We then present descriptive formulas for three families of Cantor sets, showing that the middle-third Cantor set belongs to the intersection of all three and thus serves as

a unifying example. Finally, we discuss current challenges, formulate open problems, and outline directions for future research. An appendix collects intermediate background, and interim results, and provides annotated computer code that automates some of the constructions, offering a platform for subsequent mathematical investigations.

SET-THEORETIC FOUNDATIONS

This section deals with the set theoretic foundations for description of Cantor sets. We start with fundamental definitions of F_σ, G_δ and the Borel Hierarchy (Moschovakis 2009; Kechris 1995). Then, we present the proposed set theoretic representations of Cantor sets in hierarchy from macro-level to micro-level in four levels: (i) The general representation, (ii) the nested representation, (iii) the Iterated Function System representation, and, (iv) the $q-$ ary expansion. Recurring specialized terms are collected in Appendix C for the convenience of non-specialists.

The F_σ and G_δ Properties and the Borel Hierarchy

Let $\mathcal{B}([0,1])$ denote the Borel σ-algebra on the closed unit interval. Set

$$\Sigma_1^0 := \{\, U \subseteq [0,1] \mid U \text{ is open}\},$$
$$\Pi_1^0 := \{\, F \subseteq [0,1] \mid F^c \in \Sigma_1^0\} \ (= \text{ all closed sets}),$$

and define the *Borel hierarchy* inductively for $n \geq 1$ by:

$$\Sigma_{n+1}^0 := \left\{ \bigcup_{k=1}^{\infty} A_k \;\middle|\; A_k \in \Pi_n^0 \right\},$$
$$\Pi_{n+1}^0 := \left\{ \bigcap_{k=1}^{\infty} B_k \;\middle|\; B_k \in \Sigma_n^0 \right\}.$$

The classes Σ_2^0 and Π_2^0 are traditionally denoted:

$$F_\sigma := \Sigma_2^0 \quad (\textit{countable unions of closed sets}),$$
$$G_\delta := \Pi_2^0 \quad (\textit{countable intersections of open sets}).$$

Thus, every closed subset of $[0,1]$ is F_σ, and every open subset of $[0,1]$ is G_δ. Also, the middle-third Cantor set is simultaneously F_σ and G_δ (Kechris 1995).

General Representation

At the broadest level, a Cantor set is identified with whatever closed remainder is left in $[0,1]$ after excising a countable family of pairwise disjoint open intervals; no further constraint is placed on how those intervals are chosen, distributed, or scaled. This representation therefore captures the largest class of objects that the term "Cantor set" can reasonably denote on the real line. A precise set-theoretic formulation of this viewpoint was given by Astels (1999). To begin with, let $I_0 = [0,1]$ be the closed unit interval and let $\{I_n^*\}_{n\geq 1}$ be a (finite or countable) family of pairwise–disjoint open sub-intervals contained in I_0. The general definition of Cantor set Γ is given by:

$$\Gamma = I_0 - \bigcup_{n=1}^{+\infty} I_n^*. \tag{1}$$

Considering closed sets $I_n = I_0 - I_n^* (n \in \mathbb{N})$ we have the following equivalent for the most general representation of Cantor set:

$$\Gamma = \bigcap_{n=0}^{+\infty} I_n. \tag{2}$$

This definition covers many non-fractal subsets of the closed unit interval. Furthermore, by definition Γ is closed subset of $[0,1]$ and hence a F_σ set. Also, given that for the Euclidean distance d_E we have $\Gamma = \cap_{n=1}^{+\infty} \Gamma^{(n)}$ where $\Gamma^{(n)} := \{x \in [0,1] | d_E(x,\Gamma) < \frac{1}{n}\} (n \in \mathbb{N})$ are open sets, it follows that Γ is a G_δ set.

Nested Representation

The nested representation refines the general one by demanding that the construction proceed in stages: at every step, the surviving set is a finite union of closed intervals, each of which is subdivided in the next step, with diameters tending uniformly to zero. This staged structure, absent from the general representation, is what enables stage-by-stage geometric and measure-theoretic analysis. A formal version of this more structured viewpoint was introduced by Falconer (2014) where a nested structure is imposed on their construction process as follows. Let $(s_n)_{n\geq 0}$ be a strictly increasing sequence of natural numbers with $s_0 = 1$ and $s_n \to \infty$. Construct closed sets $I_n \subset [0,1]$ for given $n \geq 0$ recursively as follows:

1. Base step. Set $I_0 = [0,1]$ and denote its single component by $I_{0,1}$.
2. Inductive step. Assume $I_{n-1} = \bigcup_{k=1}^{s_{n-1}} I_{n-1,k}$ is the disjoint union of s_{n-1} closed intervals. Subdivide each $I_{n-1,k}$ into $m_{n,k} \geq 1$ pairwise–disjoint closed sub-intervals. From the entire collection of these sub-intervals choose exactly s_n of them (at least one descendant from every parent), relabel them $I_{n,1}, \ldots, I_{n,s_n}$, and set:

$$I_n \;=\; \bigcup_{k=1}^{s_n} I_{n,k} \subseteq I_{n-1} \ \ (n \in \mathbb{N}). \tag{3}$$

Assume moreover that the intervals shrink, i.e. $\max_{1\leq k\leq s_n} |I_{n,k}| \downarrow 0$. Then $I_0 \supset I_1 \supset I_2 \supset \cdots$, with $I_n = \bigcup_{k=1}^{s_n} I_{n,k}$ and $s_n \uparrow +\infty$. The associated Cantor-set limit set is:

$$\Gamma \;=\; \bigcap_{n=0}^{+\infty} I_n. \tag{4}$$

This definition encompasses a broad class of real-line fractals, characterized by their symmetry, Lebesgue measure, and key topological attributes, most notably perfectness and nowhere denseness.

Iterated Function System (IFS) Representation

The IFS representation further specializes the nested construction by requiring that the surviving sub-intervals at every stage arise as images of $[0,1]$ under a fixed finite family of affine contractions. The resulting Cantor set is then the unique non-empty compact attractor of the system, and its self-similarity becomes a structural identity rather than an asymptotic property. This more rigid formulation is naturally expressed in the language of iterated-function systems, and appears as a special instance of the nested representation in (Falconer 2014). In this case, at each stage of the construction the surviving sub-intervals arise as images of the unit interval under a fixed finite family of affine maps. To see construction, we recall the relevant terminology as follows. First, we consider the following special maps:

1. A map $T : [0,1] \to [0,1]$ is *affine* if $T(tx_1 + (1-t)x_2) = tT(x_1) + (1-t)T(x_2)$ for all $t, x_1, x_2 \in [0,1]$.
2. It is a *contraction* if $|T(x_1) - T(x_2)| \leq r|x_1 - x_2|$ for some $r \in (0,1)$ and all $x_1, x_2 \in [0,1]$.

Second, every affine contraction can be written $T(x) = rx + s$ with $r, s \in [0,1]$ and $r < 1$ (Edgar 2007); we call $r = \text{cont. coeff}(T)$ its *contraction coefficient*. Third, let $\mathcal{T} = \{T_k\}_{k=1}^{K}$ $(K \geq 2)$ be a finite family of affine contractions with $\max_{1\leq k\leq K} \text{cont. coeff}(T_k) < 1$. Such a family constitutes an IFS. For a non-empty compact set $I \subset [0,1]$ define:

$$T(I) \quad = \quad \bigcup_{k=1}^{K} T_k(I). \tag{5}$$

The operator T is itself a contraction with ratio $r = \max_{1\leq k\leq K} \text{cont. coeff}(T_k)$ (Falconer 2014).

Hence the sequence $\{T^{\circ n}([0,1])\}_{n=0}^{+\infty}$, where $T^{\circ n}$ denotes the n-fold composition and $T^{\circ 0} = \text{id}$, converges in the Hausdorff metric to a unique non-empty compact set, called the *attractor* of the system, and we set the Cantor set as:

$$\Gamma \quad = \quad \bigcap_{n=0}^{+\infty} T^{\circ n}([0,1]). \tag{6}$$

Here, each $T^{\circ n}([0,1])$ is the disjoint union of K^n closed intervals of length at most r^n, so their diameters tend to 0 as $n \longrightarrow +\infty$. Furthermore, the Cantor set satisfies the invariance relation:

$$\Gamma \quad = \quad T(\Gamma) \; = \; \bigcup_{k=1}^{K} T_k(\Gamma). \tag{7}$$

Thus, the affine–IFS description furnishes the more specific and informative representation in the nested hierarchy of set-theoretic representations of Cantor sets.

q-Ary Expansion

The q–ary expansion is the most rigid level of the hierarchy: every affine contraction in the IFS is required to share the common slope $1/q$, so each retained interval at stage n has length exactly q^{-n} and each point of the limit set admits a digit-string address in $\mathcal{A}^{\mathbb{N}}$. This arithmetization is what makes explicit closed-form descriptions possible. The $q-$Ary expansion of Cantor sets is indeed an special case of their IFS representation when the involved affine maps have equal slopes. In details, let $q \in \mathbb{N}+2$, and $A \subset \{0,1,\cdots,q-1\}, |A| = K (2 \leq K \leq q)$. For every digit $a \in A$ define the a^{th} strict contraction T_a with slope $\frac{1}{q}$ and y-intercept $\frac{a}{q}$:

$$\begin{aligned} T_a \quad &: \quad [0,1] \longrightarrow [0,1] \\ T_a(x) \quad &= \quad \frac{1}{q}x + \frac{a}{q}. \end{aligned} \tag{8}$$

Now, the IFS representation of the Cantor set Γ is given by:

$$\Gamma \quad = \quad \bigcap_{n=0}^{+\infty} I_n : I_0 = [0,1], I_n = \bigcup_{a\in A} T_a(I_{n-1}), \;\; (n \in \mathbb{N}). \tag{9}$$

This special case has the following properties:

1. Each component of I_n is a closed interval of length q^{-n}.
2. There are K^n such components with one for each finite digit block $\vec{a} = (a_1, \cdots, a_n) : a_k \in A (1 \leq k \leq n)$.
3. The system is nested: $I_n \supset I_{n-1} (n \in \mathbb{N})$.

A straightforward verification by mathematical induction shows that:

$$\begin{aligned} I_n \quad &= \quad T_{a_1} \circ \cdots \circ T_{a_n}([0,1]) \\ &= \quad [\sum_{k=1}^{n} \frac{a_k}{q^k}, \sum_{k=1}^{n} \frac{a_k}{q^k} + \frac{1}{q^n}], (n \in \mathbb{N}). \end{aligned} \tag{10}$$

Also, the map $F_n = T_{a_1} \circ \cdots \circ T_{a_n}$ is q^{-n}-Lipschitz continuous $(n \in \mathbb{N})$ (Searcoid 2006). Consequently, $x \in \Gamma$ if and only if there is a unique (up to the standard endpoint ambiguity) infinite digit string $\vec{a}^* = (a_1, a_2, a_3, \cdots) \in A^{\mathbb{N}}$ such that $x = \sum_{n=1}^{+\infty} \frac{a_n}{q^n}$. Hence, we have the $q-$ ary expansion of the Cantor set:

$$\Gamma \quad = \quad \{x \in [0,1] | x = \sum_{n=1}^{+\infty} \frac{a_k}{q^n}, a_n \in A\}. \tag{11}$$

Notation: Henceforth, we denote $\Gamma(\alpha, K)$ middle $\alpha-$ Cantor set defined by K affine maps.

In the upcoming subsections, we study, explicit and recursive representation of three families of Cantor sets given their associated $q-$ary expansions.

Hierarchical Relationships and Practical Differences

The four representations introduced in the previous four sections form a strict descending chain of structural specificity:

$$\text{General} \supsetneq \text{Nested} \supsetneq \text{IFS} \supsetneq q\text{–ary}. \tag{12}$$

Each inclusion is proper, and at each step a concrete structural commitment is added to the preceding level:

- *General → Nested.* The arbitrary countable family of deleted open intervals is reorganized into a stagewise construction in which the surviving set at each stage is a finite union of closed intervals with diameters tending to zero. The strictness of the inclusion follows from the existence of closed nowhere-dense subsets of $[0,1]$ that admit a representation of the form (1) but not a stagewise nested decomposition with $s_n \uparrow \infty$ (see (Astels 1999)).
- *Nested → IFS.* The stagewise construction is required to be generated by a fixed finite family of affine contractions, so that the limit set satisfies the self-similarity identity $\Gamma = \bigcup_{k=1}^{K} T_k(\Gamma)$. Not every nested Cantor set is an IFS attractor: nested constructions in which the contraction ratios vary across stages or across components fail this identity (cf. (Hutchinson 1981; Falconer 2014)).
- *IFS → q–ary.* All affine contractions are forced to share the common slope $1/q$ for some integer $q \geq 2$. IFS attractors with unequal contraction ratios—e.g. the standard examples with $r_1 \neq r_2$ used to realize prescribed Hausdorff dimensions—are excluded.

The price of descending the hierarchy is generality; the reward is descriptive power. Table 1 summarizes the practical trade-offs.

The three families treated in Section illustrate the lower end of this hierarchy concretely: Γ_1 and Γ_2 are presented through their q–ary expansions and admit closed-form gap descriptions, while $\Gamma_3(p/q, 2)$, although IFS-generated, requires the recursive endpoint treatment of Theorem 0.3 because its retained-interval lengths do not collapse to a single power of a fixed base.

Table 1 Practical comparison of the four set-theoretic representations of Cantor Sets on the real line

#	Item	General	Nested	IFS	q–ary
1	Operative Parameter	Family $\{I_n^*\}$ of deleted open intervals	Sequence (s_n) and per-stage subdivisions	Affine maps $\{T_k\}_{k=1}^K$	Base q and digit set $\mathcal{A} \subseteq \{0,\ldots,q-1\}$
2	Stage-n Structure	Not stage-defined in general	s_n closed intervals of varying length	K^n closed intervals, length $\leq r^n$	K^n closed intervals of length exactly q^{-n}
3	Self-Similarity	None required	None required	Exact: $\Gamma = \bigcup_k T_k(\Gamma)$	Exact, with uniform ratio $1/q$
4	Closed-form Points	Not available in general	Not available in general	Available via fixed-point address	Available via digit expansion $\sum a_n q^{-n}$
5	Typical Use	Existence and Borel-class arguments	Stagewise measure / dimension analysis	Self-similarity, dimension theorems	Explicit formulas, arithmetic structure

DESCRIPTIVE FORMULAS FOR CANTOR SETS

Thin Family 1

The earliest set-theoretic explicit description of Cantor sets appears in (Soltanifar 2006b), where the author treats the symmetric thin family generated by at least two affine maps in an iterated-function system. This construction yields a fully constructive proof of the Hausdorff-Dimension Theorem (which asserts that for every $s > 0$ there exists a continuum of compact subsets of $\mathbb{R}^n \quad (n \geq \lceil s \rceil)$ with Hausdorff dimension of exactly s), thereby establishing the existence of the corresponding fractals for any prescribed Hausdorff dimension. Subsequent work extends the argument to broader deterministic and stochastic settings (Soltanifar 2021, 2022).

Theorem 0.1 *Let* $\Gamma_1(\frac{1}{q}, \frac{q+1}{2}) := \{x \in [0,1] | x = \sum_{n=1}^{+\infty} \frac{a_n}{q^n}\ ,\ a_n = 0, 2, 4, \cdots, q-1\}$ *where* $q \in 2\mathbb{N}+1$. *Then:*

$$\Gamma_1(\frac{1}{q}, \frac{q+1}{2}) \quad = \quad [0,1] - \bigcup_{n=1}^{+\infty} \bigcup_{k=0}^{q^{n-1}-1} \bigcup_{r=1}^{\frac{q-1}{2}} (\frac{qk+(2r-1)}{q^n}, \frac{qk+2r}{q^n}) : \quad (q \in 2\mathbb{N}+1). \quad (13)$$

In particular, for the case $q = 3$ we obtain the explicit formulae for the middle-third Cantor set $C = \Gamma_1(\frac{1}{3}, 2)$ as in the second formulae in Corollary 0.4.

Thin Family 2

In the same year, the explicit set-theoretic description of a second family of Cantor sets was presented (Soltanifar 2006a). As with the earlier construction, the sets obtained are thin and symmetric; however, their underlying iterated-function system involves only two affine contractions, in contrast to generally larger number used for the first family.

Theorem 0.2 *Let* $\Gamma_2(\frac{q-2}{q}, 2) := \{x \in [0,1] | x = \sum_{n=1}^{+\infty} \frac{a_n}{q^n}\ ,\ a_n = 0, q-1\}$ *where* $q \in \mathbb{N}+2$. *Then:*

$$\Gamma_2(\frac{q-2}{q}, 2) \quad = \quad [0,1] - \bigcup_{n=1}^{+\infty} \bigcup_{k=0}^{q^{n-1}-1} (\frac{qk+1}{q^n}, \frac{qk+2}{q^n}) : \quad (q \in \mathbb{N}+2). \quad (14)$$

Similar to the Thin Family Γ_1, for the case $q = 3$ we obtain the explicit formulae for the middle-third Cantor set $C = \Gamma_2(\frac{1}{3}, 2)$ as in the second formulae in Corollary 0.4.

Augmented Thick Family 1

In both preceding cases we provided closed-form descriptions for two families of measure-zero Cantor sets. This naturally raises the question: Does an analogous explicit representation exist for Cantor sets of positive Lebesgue measure? Current evidence suggests a negative answer; no such closed formula is known at the time of writing. Nevertheless, a constant-recursive representation remains feasible (Elaydi 2005), and the requisite background together with interim results is presented in Appendix A for interested readers.

Theorem 0.3 *Let* $\Gamma_3(\frac{p}{q}, 2) := \{x \in [0,1] | x = \sum_{n=1}^{+\infty} \frac{a_n}{(2q)^n}\ ,\ a_n = 0, 1, \cdots, q-p-1, q+p, \cdots, 2q-1,\}$ *where* $p, q \in \mathbb{N}, (p,q) = 1, 0 < \frac{p}{q} \leq \frac{1}{3}$. *Then:*

$$\begin{aligned}
\Gamma_3(\frac{p}{q}, 2) \quad = \quad & \bigcap_{n=0}^{+\infty} \bigcup_{k=1}^{2^n} [a_{n,k}(\frac{p}{q}), b_{n,k}(\frac{p}{q})] : \quad (15)\\
& a_{0,1}(\frac{p}{q}) = 0;\\
& a_{n,k}(\frac{p}{q}) = 1_{odd}(k)\Big(a_{n-1,\frac{k+1}{2}}(\frac{p}{q})\Big)\\
& +1_{even}(k)\Big(a_{n-1,\frac{k}{2}}(\frac{p}{q}) + \frac{(1-3\frac{p}{q})+(1-\frac{p}{q})(2\frac{p}{q})^n}{(1-2\frac{p}{q})2^n}\Big);\\
& b_{0,1}(\frac{p}{q}) = 1;\\
& b_{n,k}(\frac{p}{q}) = 1_{even}(k)\Big(b_{n-1,\frac{k}{2}}(\frac{p}{q})\Big)\\
& +1_{odd}(k)\Big(b_{n-1,\frac{k+1}{2}}(\frac{p}{q}) - \frac{(1-3\frac{p}{q})+(1-\frac{p}{q})(2\frac{p}{q})^n}{(1-2\frac{p}{q})2^n}\Big) :\\
& 1 \leq k \leq 2^n, n \in \mathbb{N}.
\end{aligned}$$

The Computer software code to calculate the end points $a_{n,k}(.), b_{n,k}(.)$ are given in Appendix B with associated examples

for $\alpha_1 = \frac{1}{3}$, and $\alpha_2 = \frac{1}{4}$ (R Core Team 2022). It is noteworthy to mention that the middle-third Cantor set belongs to all three discussed families. Figure 1 summarizes our above discussions in terms of hierarchy of descriptive representations.

Finally, for the case $p = 1, q = 3$ we obtain the first order recursive formulae for the middle-third Cantor set $C = \Gamma_3(\frac{1}{3}, 2)$ as in the third formulae in the following Corollary 0.4:

Corollary 0.4 *The middle-third Cantor set has three equivalent set theoretic descriptive representations:*

$$C = \{x \in [0,1] | x = \sum_{n=1}^{+\infty} \frac{a_n}{3^n},\ a_n = 0, 2\} \tag{16}$$

$$= [0,1] - \bigcup_{n=1}^{+\infty} \bigcup_{k=0}^{3^{n-1}-1} (\frac{3k+1}{3^n}, \frac{3k+2}{3^n}) \tag{17}$$

$$= \bigcap_{n=0}^{+\infty} \bigcup_{k=1}^{2^n} [a_{n,k}, b_{n,k}]: \tag{18}$$

$$a_{0,1} = 0; a_{n,k} = 1_{odd}(k)\left(a_{n-1,\frac{k+1}{2}}\right) + 1_{even}(k)\left(a_{n-1,\frac{k}{2}} + \frac{2}{3^n}\right);$$

$$b_{0,1} = 1; b_{n,k} = 1_{odd}(k)\left(b_{n-1,\frac{k+1}{2}} - \frac{2}{3^n}\right) + 1_{even}(k)\left(b_{n-1,\frac{k}{2}}\right);$$

$$1 \leq k \leq 2^n, n \in \mathbb{N}.$$

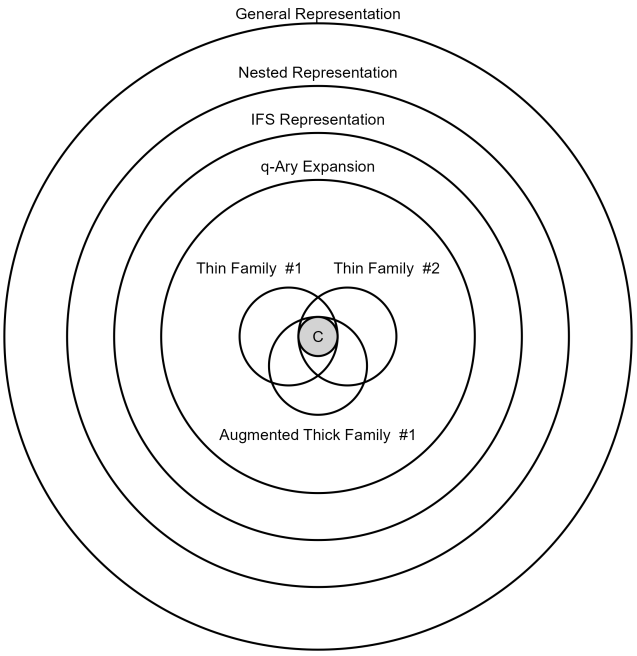


Figure 1 Diagram of descriptive formulas for the Cantor sets from set-theoretic perspective: C refers to the middle-third Cantor set

DISCUSSION

The four descriptions surveyed here, general, nested, IFS, and q–ary, form a strict hierarchy in which each level imposes additional structure on its predecessor, and the three families $\Gamma_1, \Gamma_2, \Gamma_3$ populate this hierarchy at successively finer levels of specificity. The recursive endpoint formula for $\Gamma_3(p/q, 2)$ closes a gap in the positive-measure case: where the thin families admit closed-form gap descriptions, the augmented thick family had previously resisted such an explicit treatment, and the recursion presented in Theorem 0.3 now makes every retained interval at every finite stage exactly computable. The middle-third Cantor set, sitting in the intersection of all three families, illustrates how a single canonical object can serve as a stress test for the entire descriptive framework. The remainder of this section explores three consequences of this synthesis: a continuum-theoretic reading of Γ_3 via embeddings into standard continua, a measure-theoretic construction of canonical self-similar measures supported on $\Gamma_3(\alpha)$, and an outlook tying the framework to the open problems of Section **??**.

Continuum-theoretic perspective: Although Cantor sets are totally disconnected, they occur naturally inside many *continua* (compact connected metric spaces) as closed, perfect, nowhere-dense subsets (Nadler 1992). The explicit endpoint recursion for Γ_3 supplies fine control of gap geometry at every stage. This facilitates: (i) explicit embeddings of $\Gamma_3(\alpha)$ into standard continua (e.g., as closed subsets of *arcs* (homeomorphic images of $[0,1]$), *dendrites* (locally connected continua containing no simple closed curve), or *Peano continua* (locally connected continua, equivalently continuous images of $[0,1]$)) with prescribed gap patterns; (ii) constructive "fill-in" procedures that yield continua whose complements are modeled by a chosen $\Gamma_3(\alpha)$; and (iii) numerical investigation of intersection phenomena (useful in studying when two embedded Cantor-type sets intersect), where computable gap data and thickness proxies are central. In this sense the paper's formulas act as a bridge, turning qualitative continuum-theory questions into explicit, checkable computations.

Measure-theoretic perspective (a canonical Γ_3 measure): For each $\Gamma_3(\alpha)$ we can *canonically* equip the set with a Borel probability measure that generalizes the Cantor–Lebesgue measure (Rosenthal 2006). At stage n there are q^n retained intervals; assign either uniform weights q^{-n} or, more generally, a fixed weight vector $w = (w_1, \ldots, w_q)$ with $\sum_{i=1}^{n} w_i = 1$, propagated multiplicatively along the recursion. This produces a self-similar probability measure $\mu_{\alpha,w}$ supported on $\Gamma_3(\alpha)$ whose cumulative distribution function is a Devil's–staircase generalization of the Cantor function: it is continuous, nondecreasing, constant on deleted gaps, and increases exactly on $\Gamma_3(\alpha)$. The explicit endpoints $(a_{n,k}, b_{n,k})$ allow fast, exact evaluation of cylinder-set masses and numerical plots of the associated distribution function. This construction is immediate from the framework developed here and is implemented by minor extensions of Appendix B.

Because the geometry of Γ_3 is explicit at every finite stage, the framework supports systematic computation of gap-length statistics, empirical thickness proxies, and intersection behavior under translation, tools that are relevant both to continuum-theoretic embeddings and to questions in additive/fractal geometry. We conclude with open problems on asymmetry, irrational endpoints, variable-gap deletions, the role of symmetry, and common intersections; these problems are now posed in a setting where algorithmic experiments and exact finite-stage calculations are straightforward.

The preceding survey has shown that three natural families of deterministic Cantor sets, two of measure zero with explicit descriptions and one of non-negative measure defined recursively, share a common symmetric architecture and possess rational construction end-points; indeed, the classical middle-third Cantor set lies in the intersection of all three. These parallel features highlight several directions in which the present set-theoretic framework remains incomplete and invite further research. We organize the open problems into three thematic clusters reflecting their logical

interdependence: *structural generalization* (Theme A), *variable-scale constructions* (Theme B), and *interaction phenomena* (Theme C), with the last depending on progress in the first two.

Theme A (Structural Generalization)

The descriptive formulas of Section rely on two structural restrictions—bilateral symmetry of the deletion pattern and rationality of the retained-interval endpoints. The following two problems ask whether either restriction can be dropped without losing explicitness.

OP1: *Asymmetric constructions.* Devise explicit or recursive set-theoretic formulas for Cantor sets generated by iterated-function systems with $K \geq 2$ maps that do not preserve bilateral symmetry.

OP2: *Irrational endpoints.* Find descriptive formulas when the retained intervals have irrational endpoints, again allowing $K \geq 2$ IFS components.

Rationale and impact: Symmetric, rationally-located Cantor sets form a measure-zero subfamily within the space of all IFS attractors, yet they are essentially the only ones for which closed or recursive descriptive formulas are currently known. A resolution of OP1 would extend the techniques of Section to the asymmetric self-similar sets that arise as hyperbolic invariant sets in low-dimensional dynamics (Palis and Takens 1993; Moreira and Yoccoz 2001), where asymmetry is the rule rather than the exception. A resolution of OP2 would connect the descriptive framework to the algebraic theory of self-similar measures with non-rational contraction ratios, where deep recent work by Hochman (Hochman 2014) and Shmerkin (Shmerkin 2019) has clarified the dimension theory but explicit set-theoretic formulas remain absent.

Tools and feasibility: The endpoint recursion of Theorem 0.3 is candidate machinery for OP1: removing the symmetry assumption requires tracking left- and right-endpoints under independent contraction ratios $r_L \neq r_R$, which extends the recursion in Lemmas 4.1–4.3 (Appendix A) at the cost of a more elaborate index scheme. OP2 is genuinely harder, since the arithmetic identity $p/q = 2p/(2q)$ that drives the $2q$-ary expansion in Appendix A has no analogue for irrational ratios; symbolic-dynamic coordinates (Lind and Marcus 1995) or continued-fraction expansions of the endpoints offer alternative descriptive routes that have not yet been systematically explored.

Theme B (Variable-Scale Constructions)

The middle-α family $\Gamma_3(\alpha, 2)$ removes a gap of length α^n at stage n, so the deletion sequence is geometric. The following two problems concern non-geometric deletion schemes and the role of symmetry within them.

OP3: *Variable-gap deletions.* Given a sequence $(\alpha_n)_{n\geq 1}$ with $0 < \alpha_n < 1$ satisfying $1 - \sum_{n=1}^{\infty} 2^{n-1}\alpha_n \geq 0$, provide a recursive descriptive formula for the Cantor set obtained by removing at stage n the open middle interval of length α_n.

OP4: *Role of symmetry.* Determine how the solution of OP3 varies when bilateral symmetry of the deletion is enforced or relaxed.

Rationale and impact: Variable-scale constructions interpolate between the self-similar Cantor sets of Section and the broader nested representations of (4). Smith–Volterra–Cantor sets and their generalizations belong to this class (Vallin 2013), and their Lebesgue measure can be tuned arbitrarily within $[0, 1)$ by choice of (α_n). A resolution of OP3 would yield exact endpoint formulas for these positive-measure constructions, parallel to what Theorem 0.3 achieves in the self-similar case, and would supply concrete examples for testing fractal-uncertainty principles in the variable-scale setting (Łaba and Pramanik 2009). OP4 is significant because asymmetric variable-gap constructions exhibit qualitatively different harmonic-analytic behavior—in particular, the Fourier decay of the associated self-similar measures depends sensitively on whether deletions are symmetric.

Tools and feasibility: The recursion of Lemma 4.3 generalizes directly to OP3 by replacing α^n with α_n in the increment $\Delta_{n,k}$; the closed form of $\Delta_{n,k}$ no longer collapses but remains explicit as a finite sum at each stage. OP4 admits a tractable first case (left/right asymmetry only, with all other parameters symmetric), which would be a natural starting point.

Theme C (Interaction Phenomena)

Themes A and B describe families of Cantor sets indexed by structural parameters (asymmetry, irrationality, deletion sequence). The next problem asks how families from these themes interact.

OP5: *Common ground.* Establish conditions under which the families arising in OP1–OP4 admit a non-empty mutual intersection, and characterize the intersection when it exists.

Rationale and impact: The middle-third Cantor set is, in the present survey, the unique explicit point of intersection across the three families $\Gamma_1, \Gamma_2, \Gamma_3$. Whether analogous canonical objects exist across the broader families of OP1–OP4 is a question of both descriptive and dynamical interest. Common intersections are the deterministic shadow of the Newhouse phenomenon and the Palis conjecture (Newhouse 1979; Palis and Takens 1993; Moreira and Yoccoz 2001): when do two Cantor sets meet, and when does the meeting persist under perturbation? A resolution of OP5 would supply concrete test cases for these conjectures and would bridge the explicit set-theoretic framework of this survey to the geometric measure-theoretic literature on intersections (Mattila 1995).

Tools and feasibility: The thickness theory of Astels (Astels 1999, 2000) provides sufficient conditions for non-empty intersection of two Cantor sets in terms of their gap structures, and the explicit endpoint data supplied by Theorem 0.3 (and its Theme-A and Theme-B extensions) make these thickness computations exact rather than asymptotic. Moreira–Yoccoz (Moreira and Yoccoz 2001) provides the deeper machinery for *stable* intersections under perturbation, which becomes relevant once OP1–OP4 have been resolved.

Resolving these questions would extend the taxonomy of Cantor sets beyond the symmetric, rational, geometrically-scaled cases treated here and clarify how set-theoretic representations interact with geometric asymmetry, irrational numbers, variable deletion schemes, and the intersection phenomena that link them. The thematic grouping above is intended to make explicit the logical dependence, Theme C builds on Themes A and B, and to facilitate parallel progress on the structural and variable-scale fronts.

CONCLUSION

This survey brings four set-theoretic perspectives, general, nested, IFS, and $q-$ary, under one framework and demonstrates how they meet in three representative Cantor-set families containing the middle-third classic. The synthesis clarifies which descriptive formulas are now complete and which remain missing, especially for asymmetric, irrational-endpoint, and variable-gap constructions.

By distilling these gaps into a concise list of open problems, the paper sets a clear agenda for future work in descriptive set theory and fractal analysis on the real line.

Abbreviations

The following abbreviations are used in this manuscript: ACDC: Applications and Cross-Disciplinary Connections; IFS: Iterated Function System; NC: Novelty and Contribution; OP: Open Problem; SVC: Smith–Volterra–Cantor.

Acknowledgments

The author is grateful to the reviewers for their constructive comments, which significantly improved the presentation of this manuscript.

Ethical standard

The author has no relevant financial or non-financial interests to disclose.

Availability of data and material

Not applicable.

Conflicts of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding

This research received no external funding.

APPENDIX

Appendix A: Interim Results

Let $\Gamma_3(\alpha,2)$ where $0<\alpha\leq\frac{1}{3}$ be the middle $-\alpha$ Cantor set defined by two affine maps. Here, its nested representation is given by $\Gamma_3(\alpha,2)=\cap_{n=0}^{+\infty}I_n(\alpha)$ where $I_0(\alpha)=[0,1]$ and $I_n(\alpha)$ is defined recursively by removing 2^{n-1} middle open intervals each of the length α^n from $I_{n-1}(\alpha)$ $(n\in\mathbb{N})$. We have the following interim results:

1. **Family Name** By construction, the Lebesgue measure is given by $\lambda(\Gamma_3(\alpha,2))=1-\sum_{n=1}^{+\infty}2^{n-1}\alpha^n=\frac{1-3\alpha}{1-2\alpha}(0<\alpha\leq\frac{1}{3})$. Hence, we augment the "Thick Family" with $0<\alpha<\frac{1}{3}$ with the "Thin case" $\alpha=\frac{1}{3}$; and, call it the "Augmented Thick Family" with $0<\alpha\leq\frac{1}{3}$.

2. **2q-Ary Expansion** When $0<\alpha=\frac{p}{q}\leq\frac{1}{3}, p,q\in\mathbb{N},(p,q)=1$, by above construction, the equality $\frac{p}{q}=\frac{2p}{2q}$, and mathematical induction, a straightforward verification shows the following 2q-Ary expansion:

$$\begin{aligned}\Gamma_3(\alpha,2) &= \{x\in[0,1]|x=\sum_{n=1}^{+\infty}\frac{a_n}{(2q)^n},\\ &\quad a_n=0,1,\cdots,q-p-1,q+p,\cdots,2q-1\}.\end{aligned}$$

   This representation facilitates the n-Aray expansion for cases with even q. Furthermore, the middle-third Cantor set has the secondary 6-Ary expansion as $C=\{x\in[0,1]|x=\sum_{n=1}^{+\infty}\frac{a_n}{6^n},\ a_n=0,1,4,5\}$.

3. **Proof of Theorem 0.3** We prove the Theorem in steps (a)-(e).

(a) **Notation** Given above recursive construction it follows that, $\Gamma_3(\alpha,2)=\cap_{n=0}^{+\infty}I_n(\alpha)$ where $I_n(\alpha)=\cup_{k=1}^{2^n}I_{n,k}(\alpha)$ and $I_{n,k}(\alpha)=[a_{n,k}(\alpha),b_{n,k}(\alpha)],(1\leq k\leq 2^n,n\in\mathbb{N}_0)$.

(b) **Lemma 0.5** $\lambda(I_{n,k}(\alpha))=\frac{(2\alpha)^{n+1}+2(1-3\alpha)}{2^{n+1}(1-2\alpha)}:\ (1\leq k\leq 2^n,n\in\mathbb{N}_0)$.

Proof. The case for $n=0$ is trivial. Let $n\in\mathbb{N}$ and $J_{n,k}(\alpha)$ be the middle open interval of length α^n removed from $I_{n-1,k}(\alpha)$ in the nested construction process i.e. $I_{n-1,k}(\alpha)=I_{n,2k-1}(\alpha)\cup J_{n,k}(\alpha)\cup I_{n,2k}(\alpha)$. Then, given, $\lambda(I_{n,2k-1}(\alpha))=\lambda(I_{n,2k}(\alpha))$ we have: $\lambda(I_{n-1,k}(\alpha))=2\lambda(I_{n,k}(\alpha))+\alpha^n$. Consequently, we have a first order recursive relation:

$$\lambda(I_{n,k}(\alpha))=\frac{1}{2}\lambda(I_{n-1,k}(\alpha))-\frac{\alpha^n}{2},\lambda(I_{0,1}(\alpha))=1.$$

Finally, the assertion follows by mathematical induction on $n\in\mathbb{N}$.

(c) **Lemma 0.6** *Let* $\Delta_{n,k}(\alpha)=\lambda(I_{n,k}(\alpha))+\alpha^n\ \ (1\leq k\leq 2^n,n\in\mathbb{N}_0)$. *Then,*

$$\Delta_{n,k}(\alpha)=\frac{(1-3\alpha)+(1-\alpha)(2\alpha)^n}{(1-2\alpha)2^n}\ (1\leq k\leq 2^n,n\in\mathbb{N}_0).$$

Proof. This is straightforward result from Lemma 0.5.

(d) **Lemma 0.7** *Let* $a_{n,k}(\alpha),b_{n,k}(\alpha),(1\leq k\leq 2^n,n\in\mathbb{N}_0)$ *be defined as in above recursive construction. Then:*

$$\begin{aligned}a_{0,1}(\alpha) &= 0,\\ a_{n,k}(\alpha) &= 1_{odd}(k)\Big(a_{n-1,\frac{k+1}{2}}(\alpha)\Big)\\ &\quad +1_{even}(k)\Big(a_{n-1,\frac{k}{2}}(\alpha)+\Delta_{n,k}(\alpha)\Big),\\ b_{0,1}(\alpha) &= 1,\\ b_{n,k}(\alpha) &= 1_{even}(k)\Big(b_{n-1,\frac{k}{2}}(\alpha)\Big),\\ &\quad +1_{odd}(k)\Big(b_{n-1,\frac{k+1}{2}}(\alpha)-\Delta_{n,k}(\alpha)\Big)\\ &\qquad (1\leq k\leq 2^n,n\in\mathbb{N}_0).\end{aligned}$$

Proof. This is straightforward result from definition of $\Delta_{n,k}(\alpha)$ in Lemma 0.6, and comparing the end points of closed intervals in the following equations:

$$\begin{aligned}[a_{n-1,k}(\alpha),b_{n-1,k}(\alpha)] &= I_{n-1,k}(\alpha)\\ &= I_{n,2k-1}(\alpha)\cup J_{n,k}(\alpha)\cup I_{n,2k}(\alpha)\\ &= [a_{n,2k-1}(\alpha),b_{n,2k-1}(\alpha)]\\ &\quad\cup(b_{n,2k-1}(\alpha),a_{n,2k}(\alpha))\\ &\quad\cup[a_{n,2k}(\alpha),b_{n,2k}(\alpha)],\\ &\qquad (1\leq k\leq 2^n,\ n\in\mathbb{N}_0).\end{aligned}$$

(e) **Completion** Finally, by considering $0<\alpha=\frac{p}{q}\leq\frac{1}{3}$ in Lemma 0.6 and substituting corresponding $\Delta_{n,k}$ in the equations presented in Lemma 0.7, the proof for the Theorem 0.3 is completed.

Appendix B: R Software Code

The following function in R software code produces the end points of the intervals of the recursive representation of "Augmented Thick Family" $\Gamma_3(\frac{p}{q},2)$ for given pair $(n,k):1\leq k\leq 2^n,n\in\mathbb{N}$. It presents two examples of calculations one for the thin Cantor set and another for the thick Cantor set:

```
## ====================================================
##  Recursive evaluator for lower endpoint a_{n,k}(p/q)
## ====================================================
a_nk <- local({
  ## simple memorization (saves repeated work)
  cache <- new.env(parent = emptyenv())
  f <- function(n, k, p, q) {
    if (n == 0L) {                      # base level
      if (k != 1L) stop("For n = 0 you must take k = 1")
      return(0)
    }
    if (k < 1 || k > 2^n)
      stop("k must satisfy 1 <= k <= 2^n")
    ## build a unique key for memorization
    key <- paste(n, k, p, q, sep = "-")
    if (exists(key, envir = cache, inherits = FALSE))
      return(cache[[key]])
    r <- p / q                          # the ratio p/q
    ## recursive step
    if (k %% 2L == 1L) {                # k is odd
      val <- f(n - 1L, (k + 1L) / 2L, p, q)
    } else {                            # k is even
      prev <- f(n - 1L, k / 2L, p, q)
      term <- ((1 - 3 * r) +
      (1 - r) * (2 * r)^n) /((1 - 2 * r) * 2^n)
      val <- prev + term
    }
    cache[[key]] <- val                 # store and return
    val
  }
  f
})

## ====================================================
##  Recursive evaluator for upper endpoint  b_{n,k}(p/q)
## ====================================================

b_nk <- local({
  cache <- new.env(parent = emptyenv())  #memorization
  f <- function(n, k, p, q) {
    if (n == 0L) {                   # base case
      if (k != 1L) stop("For n = 0 you must take k = 1")
      return(1)
    }
    if (k < 1L || k > 2L^n)
      stop("k must satisfy 1 <= k <= 2^n")
    key <- paste(n, k, p, q, sep = "-")
    if (exists(key, envir = cache, inherits = FALSE))
      return(cache[[key]])
    r    <- p / q
    term <- ((1 - 3 * r) +
    (1 - r) * (2 * r)^n) /((1 - 2 * r) * 2^n)
    ## recursive step
    if (k %% 2L) {   # k  is odd
      val <- f(n - 1L, (k + 1L) %/% 2L, p, q) - term
    } else {      # k is even
      val <- f(n - 1L, k %/% 2L, p, q)
    }
    cache[[key]] <- val
    val
  }
  f
})

## ====================================================
##  Intervals Endpoints Calculator
## ====================================================
EndpointCalculator <- function(n, p, q) {
  ## ---------- base-level endpoints ----------
  a0 <- a_nk(0L, 1L, p, q)
  b0 <- b_nk(0L, 1L, p, q)
  ## keep the original "too-large" denominator
  denom <- q^(n + 1L)
  ## ---------- level-n endpoints ----------
  res1 <- sapply(1:2^n, function(k) a_nk(n, k, p, q))
  res2 <- sapply(1:2^n, function(k) b_nk(n, k, p, q))
  num1 <- c(a0 * denom, round(res1 * denom))
  num2 <- c(b0 * denom, round(res2 * denom))
  ## -------- LABELS (divide out one factor of q) --------
  display_denom <- denom / q   #  q^(n+1) / q  ->  q^n
  col_labs      <- paste0("k=", 0:2^n)
  fractions1    <- paste0(num1 / q, "/", display_denom)
  fractions2    <- paste0(num2 / q, "/", display_denom)
  ## ---------- NEW ROW ----------
  combo <- paste0("[", fractions1, ",", fractions2, "]")
  out1   <- rbind("a_(n,k)(α)"             = fractions1,
                  "b_(n,k)(α)"             = fractions2,
                  "[a_(n,k)(α),b_(n,k)(α)]" = combo)
  colnames(out1) <- col_labs
  out2<-t(out1)
  return(out2)
}

## ====================================================
##  Examples
## ====================================================
## Example (1) : Thin Cantor Set
EndpointCalculator(n=2, p=1, q=3)
#      a_(n,k)(α) b_(n,k)(α) [a_(n,k)(α),b_(n,k)(α)]
# k=0 "0/9"      "9/9"      "[0/9,9/9]"
# k=1 "0/9"      "1/9"      "[0/9,1/9]"
# k=2 "2/9"      "3/9"      "[2/9,3/9]"
# k=3 "6/9"      "7/9"      "[6/9,7/9]"
# k=4 "8/9"      "9/9"      "[8/9,9/9]"

## Example (2) : Thick Cantor Set
EndpointCalculator(n=2, p=1, q=4)
#      a_(n,k)(α) b_(n,k)(α) [a_(n,k)(α),b_(n,k)(α)]
# k=0 "0/16"     "16/16"    "[0/16,16/16]"
# k=1 "0/16"     "2.5/16"   "[0/16,2.5/16]"
# k=2 "3.5/16"   "6/16"     "[3.5/16,6/16]"
# k=3 "10/16"    "12.5/16"  "[10/16,12.5/16]"
# k=4 "13.5/16"  "16/16"    "[13.5/16,16/16]"
```

Appendix C: Glossary

This appendix includes notational conventions and key terms used throughout the survey for the convenience of readers outside descriptive set theory and fractal geometry. Specialists may skip this subsection without loss.

The terms in the paper that a general mathematical reader (e.g., an analyst or topologist) may pause fall into four clusters:

1. **Topological / set-theoretic descriptors:** *Perfect set, nowhere dense, totally disconnected, derived set, F_σ, G_δ, Borel hierarchy.*

2. **Measure-theoretic descriptors:** *Lebesgue measure, outer content, measure-zero vs. positive measure, self-similar measure, Cantor–Lebesgue function, Devil's staircase.*

3. **Fractal / dynamical terminology:** *Hausdorff dimension, attractor, IFS, contraction coefficient, affine map, thickness.*

4. **Continuum-theoretic terms:** *Continuum, arc, dendrite, Peano continuum, embedding*

The concise mathematical definitions and descriptions are as in follows:

- *Perfect set* — a closed set in which every point is a limit point of the set; equivalently, a closed set with no isolated points (Munkres 2000).
- *Nowhere dense set* — a set whose closure has empty interior; informally, a set that is "full of holes" in every sub-interval.
- *Totally disconnected set* — a set whose only connected subsets are singletons; equivalently, between any two distinct points one can interpose a gap.
- F_σ *set* — a countable union of closed sets.
- G_δ *set* — a countable intersection of open sets.
- *Borel hierarchy* — the transfinite stratification Σ^0_n, Π^0_n of Borel sets by alternating countable unions and intersections, beginning with open and closed sets (Kechris 1995).
- *Affine map on* $[0,1]$ — a function of the form $T(x) = rx + s$; a *contraction* if $|r| < 1$, with $r = \text{cont.coeff}(T)$ its *contraction coefficient*.
- *Iterated function system (IFS)* — a finite family $\{T_k\}_{k=1}^K$ of contractions; its *attractor* is the unique non-empty compact set Γ satisfying $\Gamma = \bigcup_{k=1}^K T_k(\Gamma)$ (Falconer 2014).
- *Hausdorff dimension* — the critical exponent $\dim_H(E)$ at which the s-dimensional Hausdorff measure of E transitions from $+\infty$ to 0; for self-similar Cantor sets generated by K maps with common ratio r, $\dim_H(E) = \log K / \log(1/r)$.
- *Self-similar measure* — a Borel probability measure μ on the attractor of an IFS $\{T_k\}$ satisfying $\mu = \sum_{k=1}^K w_k\,(\mu \circ T_k^{-1})$ for a probability weight vector $(w_1, \ldots, w_K)$.
- *Cantor–Lebesgue function (Devil's staircase)* — the continuous, non-decreasing function $[0,1] \to [0,1]$ that is constant on every gap of the middle-third Cantor set and increases only on the set itself.

LITERATURE CITED

Akhadkulov, H. and Y. Jiang, 2019 Volume of the hyperbolic cantor sets. Dynamical Systems **35**: 185–196.

Astels, S., 1999 Thickness measures for cantor sets. Electronic Research Announcements of the American Mathematical Society **5**: 108–111.

Astels, S., 2000 Cantor sets and numbers with restricted partial quotients. Transactions of the American Mathematical Society **352**: 133–170.

Babich, A., 1992 Scrawny cantor sets are not definable by tori. Proceedings of the American Mathematical Society **115**: 829.

Bartoszewicz, A., M. Filipczak, S. Glcab, F. Prus-Wiśniowski, and J. Swaczyna, 2017 On generating regular cantorvals connected with geometric cantor sets. arxiv - Cornell university. Pre-print.

Batakis, A. and G. Havard, 2024 Dimensions of harmonic measures on non-autonomous cantor sets. arxiv - Cornell university. Pre-print.

Batakis, A. and A. Zdunik, 2015 Hausdorff and harmonic measures on non-homogeneous cantor sets. Annales Academiae Scientiarum Fennicae Mathematica **40**: 279–303.

Bellissard, V., J. and A. Julien, 2013 Bi-lipshitz embedding of ultrametric cantor sets into euclidean spaces. arxiv - Cornell university. Pre-print.

Brouwer, L. E. J., 1910 On the structure of perfect sets of points. Proceedings of the Royal Netherlands Academy of Arts and Sciences **12**: 785–794.

Cabrelli, C., U. Darji, and U. Molter, 2011 Visible and invisible cantor sets. arxiv - Cornell university Pre-print.

Cheraghi, D. and M. Pedramfar, 2019 Hairy cantor sets. arxiv - Cornell university. Pre-print.

Cockerill, E., 2002 Multi-model cantor sets. arxiv - Cornell university. Pre-print.

Denette, E., 2016 *Minimal Cantor Sets: The Combinatorial Construction of Ergodic Families and Semi-conjugations*. Ph.D. thesis, University of Rhode Island, Open Access Dissertations. Paper 436.

Edgar, G., 2007 *Measure, Topology, and Fractal Geometry*. Springer, New York, NY, USA.

Elaydi, S., 2005 *An Introduction to Difference Equations*. Springer, New-york, NY, USA, fourth edition.

Eswarathasan, S. and X. Han, 2023 Fractal uncertainty principle for discrete cantor sets with random alphabets. Mathematical Research Letters **30**: 1657–1679.

Falconer, K., 2014 *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons, Chichester, West Sussex, UK.

Fillman, J. and S. H. Tidwell, 2023 On sums of semibounded cantor sets. Rocky Mountain Journal of Mathematics **53**: 737–754.

Fleron, J. F., 1994 A note on the history of the cantor set and cantor function. Mathematics Magazine **67**: 136–140.

Fletcher, A., D. Stoertz, and V. Vellis, 2022 Genus g cantor sets and germane julia sets. arxiv - Cornell university Pre-print.

Fletcher, A. and J. Wu, 2014 Julia sets and wild cantor sets. arxiv - Cornell university. Pre-print.

Frame, M., B. Mandelbrot, and N. Neger, 2025 Ancient egyptian columns [web page]. in fractals in african art. Department of Mathematics. Yale University, Retrieved June 15, 2025.

Garcia, I., 2011 A family of smooth cantor sets. Annales Fennici Mathematici **36**: 21–45.

Gorodetski, A. and S. Northrup, 2015 On sums of nearly affine cantor sets. arxiv - Cornell university. Pre-print.

Hare, K. and K.-S. Ng, 2015 Hausdorff and packing measures of balanced cantor sets. Real Analysis Exchange **40**: 113.

Hare, K. G. and N. Sidorov, 2024 The minkowski sum of linear cantor sets. Acta Arithmetica **212**: 173–193.

Hassan, M. K., M. Z. Hassan, and N. I. Pavel, 2018 Extensive analytical and numerical investigation of the kinetic and stochastic cantor set. arxiv - Cornell university Pre-print:0907.4837.

Hieronymi, P., 2018 A tame cantor set. Journal of the European Mathematical Society **20**: 2063–2104.

Hochman, M., 2014 On self-similar sets with overlaps and inverse theorems for entropy. Annals of Mathematics **180**: 773–822.

Honary, B., M. Pourbarat, and M. R. Velayati, 2006 On the arithmetic difference of affine cantor sets. Journal of Dynamical Systems and Geometric Theories **4**: 139–146.

Hu, M. and S. Wen, 2008 Quasisymmetrically minimal uniform cantor sets. Topology and Its Applications **155**: 515–521.

Hutchinson, J. E., 1981 Fractals and self-similarity. Indiana University Mathematics Journal **30**: 713–747.

Jin, Z. A., 2021 Cantor sets in topology, analysis, and financial markets. Department of Mathematics. University of Chicago, Retrieved June 15, 2025.

Jumha, S. H., I. O. Hamad, and A. O. Hassan, 2024 Cantor set from a nonstandard viewpoint. Polytechnic Journal **14**, Article 12.

Kainth, S. P. S., 2023 *A Comprehensive Textbook on Metric Spaces*. Springer, Singapore.
Kamalutdinov, K. and A. Tetenov, 2018 Twofold cantor sets in r. arxiv - Cornell university. Pre-print.
Kechris, A. S., 1995 *Classical Descriptive Set Theory*. Springer, New York, NY, USA.
Kong, D., 2014 On the self similarity of generalized cantor sets. SCIENTIA SINICA Mathematica **44**: 945–956.
Krushkal, V., 2018 Sticky cantor sets in rd. Journal of Topology and Analysis **10**: 477–482.
Łaba, I. and M. Pramanik, 2009 Arithmetic progressions in sets of fractional dimension. Geometric and Functional Analysis **19**: 429–456.
Lapidus, M. L. and H. Lu, 2008 Nonarchimedean cantor set and string. Journal of Fixed Point Theory and Applications **3**: 181–190.
Li, W. and Y. Yao, 2008 The pointwise densities of non-symmetric cantor sets. International Journal of Mathematics **19**: 1121–1135.
Lin, Z., Y. Qiu, and K. Wang, 2024 Number rigid determinantal point processes induced by generalized cantor sets. arxiv - Cornell university. Pre-print.
Lind, D. and B. Marcus, 1995 *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, Cambridge, UK.
Mandelbrot, B. B., 1982 *The Fractal Geometry of Nature*. W. H. Freeman, New york, NY, USA.
Mantica, G., 2015 Numerical computation of the isospectral torus of finite gap sets and of ifs cantor sets. arxiv - Cornell university. Pre-print.
Martens, M., 1994 Distortion results and invariant cantor sets of unimodal maps. Ergodic Theory and Dynamical Systems **14**: 331–349.
Mattila, P., 1995 *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, UK.
Mauldin, R. D. and S. C. Williams, 1986 Random recursive constructions: Asymptotic geometric and topological properties. Transactions of the American Mathematical Society **295**: 325–346.
McLoughlin, M. P. M. M. and C. R. Krizan, 2014 Some results on distorted cantor sets. International Journal of Mathematical Analysis **8**: 35–49.
Moreira, C. G., M. Pacifico, and S. Romaña Ibarra, 2020 Hausdorff dimension, lagrange and markov dynamical spectra for geometric lorenz attractors. Bulletin of the American Mathematical Society **57**: 269–292.
Moreira, C. G. and J.-C. Yoccoz, 2001 Stable intersections of regular cantor sets with large hausdorff dimensions. Annals of Mathematics **154**: 45–96.
Moschovakis, Y. N., 2009 *Descriptive Set Theory*, volume 155. American Mathematical Society, Santa Monica, CA, USA.
Munkres, J. R., 2000 *Topology*. Prentice Hall, second edition.
Nadler, S. B., Jr., 1992 *Continuum theory: An introduction*. Marcel Dekker.
Nakata, T., 1997 An approximation of hausdorff dimensions of generalized cookie-cutter cantor sets. Hiroshima mathematical journal **27**: 467–475.
Nassiri, M. and M. Z. Bidaki, 2025 Cantor sets in higher dimension i: Criterion for stable intersections. arxiv - Cornell university. Pre-print.
Newhouse, S. E., 1979 The abundance of wild hyperbolic sets and non-smooth stable sets for diffeomorphisms. Publications Mathématiques de l'IHÉS **50**: 101–151.
Nowakowski, P., 2021 The family of central cantor sets with packing dimension zero. Tatra Mountains Mathematical Publications **78**: 1–8.
Palis, J. and F. Takens, 1993 *Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations: Fractal Dimensions and Infinitely Many Attractors*, volume 35 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, UK.
R Core Team, 2022 R: A language and environment for statistical computing (version 4.2.0). [Computer software]. R Foundation for Statistical Computing. Vienna, Austria.
Rani, M. and S. K. Prasad, 2010 Superior cantor sets and superior devil staircases. International Journal of Artificial Life Research **1**: 78–84.
Rosenthal, J. S., 2006 *A First Look at Rigorous Probability Theory*. World Scientific Publishing, Singapore, Singapore, second edition.
Salem, R., 1951 On singular monotonic functions whose spectrum has a given hausdorff dimension. Arkiv för Matematik **1**: 353–365.
Schroeder, M. R., 2009 *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. Dover Publications, Mineola, NY, USA.
Searcoid, M., 2006 Lipschitz functions. In *Metric Spaces*, Springer undergraduate mathematics series, Springer-Verlag, New York, NY, USA.
Semmes, S., 2011 p-adic heisenberg cantor sets, 2. arxiv - Cornell university. Pre-print.
Shmerkin, P., 2019 On furstenberg's intersection conjecture, self-similar measures, and the l^q norms of convolutions. Annals of Mathematics **189**: 319–391.
Soltanifar, M., 2006a A different description of a family of middle-α cantor sets. American Journal of Undergraduate Research **5**.
Soltanifar, M., 2006b On a sequence of cantor fractals. Rose–Hulman Undergraduate Mathematics Journal **7**: 9.
Soltanifar, M., 2021 A generalization of the hausdorff dimension theorem for deterministic fractals. Mathematics **9**: 1546.
Soltanifar, M., 2022 The second generalization of the hausdorff dimension theorem for random fractals. Mathematics **10**: 706.
Soltanifar, M., 2023 A classification of elements of function space $f(\mathbb{R},\mathbb{R})$. Mathematics **11**: 3715.
Tresser, C., 1991 Fine structure of universal cantor sets. In *Instabilities and Nonequilibrium Structures III*, edited by E. Tirapegui and W. Zeller, volume 64 of *Mathematics and Its Applications*, Springer, Dordrecht, Netherlands.
Vallin, R. W., 2013 *The Elements of Cantor Sets: With Applications*. John Wiley & Sons, Hoboken, NJ, USA.
Zeng, Y., 2017 Self-similar subsets of symmetric cantor sets. Fractals **25**: 1750003.
Zhou, Q., X. Li, and Y. Li, 2019 Deformations on symbolic cantor sets and ultrametric spaces. arxiv - Cornell university. Pre-print.

How to cite this article: Soltanifar, M. Descriptions of Cantor Sets: A Set-Theoretic Survey and Open Problems. *Chaos and Fractals*, 3(2), 65-75, 2026.

ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.76-91
https://doi.org/10.69882/adba.chf.2026072

An Explainable Leakage-Audited Framework for Multi-Horizon Sepsis-3 Onset Forecasting from Complex ICU Dynamics

Mohammed Mansour *,1 **and Turker Berk Donmez** α,2

*Mechatronics Engineering Department, Sakarya University of Applied Sciences, University Road, 54050 Serdivan, Sakarya, Türkiye, αBiomedical Engineering Department, Sakarya University of Applied Sciences, University Road, 54050 Serdivan, Sakarya, Türkiye.

ABSTRACT Sepsis-3 onset forecasting remains challenging because event prevalence is low at each patient-time landmark and retrospective electronic health record (EHR) models may inadvertently exploit workflow signals that reflect clinician suspicion rather than underlying patient physiology. We present an explainable, leakage-audited, multi-horizon LightGBM framework for physiology-based Sepsis-3 onset forecasting using MIMIC-IV v3.1. Adult first ICU stays with at least 8 hours of observation were transformed into hourly physiologic trajectories, Sepsis-3 onset labels, and repeated landmark prediction examples. Separate models estimated the probability of Sepsis-3 onset within the next 4, 6, 12, and 24 hours. The final cohort comprised 64,363 ICU stays, including 42,291 forecastable stays and 6,431 forecastable Sepsis-3 cases. The deployment-oriented no-admission-year CLEAN model achieved held-out AUROCs of 0.782, 0.772, 0.750, and 0.720 for the 4-, 6-, 12-, and 24-hour prediction horizons, respectively, with patient-level bootstrap 95% confidence intervals of 0.763–0.801, 0.752–0.792, 0.729–0.769, and 0.699–0.742. A paired leakage audit demonstrated that removing vasopressor exposure, measurement-age channels, and observation-mask channels resulted in AUROC changes of less than 0.01, indicating that predictive performance was primarily driven by physiologic information rather than workflow-related artifacts. Temporal validation on the latest deidentified admission-year group yielded a 6-hour AUROC of 0.762. At the 6-hour horizon, a top-2% no-year alarm policy with a 6-hour cooldown detected 17.8% of sepsis cases within 0–6 hours, 29.3% within 0–24 hours, and 40.3% before onset overall while generating 15.5 alerts per 100 patient-days. SHAP-based explainability analyses consistently identified neurologic status, temperature, urine-output dynamics, renal markers, inflammatory variables, vital signs, and lactate-related dynamics as the dominant predictors across forecasting horizons. These findings provide a robust and explainable physiology-based framework for future studies on Sepsis-3 onset forecasting in complex critical-care systems.

KEYWORDS
Explainable AI
SHAP
Sepsis-3
MIMIC-IV
Dynamic landmark forecasting

INTRODUCTION

Sepsis remains one of the central unsolved problems in acute care because its early clinical manifestations are nonspecific, its trajectory can change over hours, and delayed treatment is associated with preventable organ dysfunction and death. Recent clinical AI studies have moved from retrospective discrimination metrics toward real workflow impact: the COMPOSER deployment study reported that an EHR-integrated deep-learning sepsis predictor was associated with improved sepsis bundle compliance and reduced in-hospital sepsis mortality in two emergency departments (Boussina *et al.* 2024). At the same time, recent systematic reviews show that sepsis prediction models differ widely in target definitions, data windows, model classes, validation strategies, and implementation maturity (Gao *et al.* 2024b; Zubair *et al.* 2025; Zhang *et al.* 2026). These observations motivate a model-development strategy that treats timing, interpretability, and alarm burden as first-class scientific endpoints rather than secondary engineering details.

Large critical-care EHR resources have made reproducible sepsis modeling possible, but they also expose the methodological fragility of retrospective clinical prediction. MIMIC-IV provides deidentified hospital and ICU data from Beth Israel Deaconess Medical Center and is now a standard benchmark for critical-care informatics (Johnson *et al.* 2023). The current MIMIC-IV v3.1 PhysioNet release adds a citable, versioned data object for reproducible research (Johnson *et al.* 2024). Because each version changes the available tables, item definitions, and timestamps that downstream pipelines consume, a sepsis forecasting study must report not only the model family and metrics but also the database version, cohort construction, label windows, and feature provenance.

The recent machine-learning literature supports the potential of EHR-based sepsis prediction, but it also shows that high apparent performance is not sufficient for clinical credibility. A systematic review of machine-learning-based early sepsis prediction using EHRs found that most studies used longitudinal data, but only a

Manuscript received: 4 May 2026,
Revised: 28 June 2026,
Accepted: 10 July 2026.

[1]mohammedmansour@subu.edu.tr (**Corresponding author**)
[2]turkerberkdonmez@yahoo.com

minority reported long prediction horizons or prospective validation, and very few made complete code available (Islam *et al.* 2023). A 2024 network meta-analysis concluded that machine-learning algorithms can be effective for sepsis prediction, while also emphasizing heterogeneity across cohorts and algorithms (Gao *et al.* 2024b). More recent reviews similarly highlight that reported AUROCs can be high while evidence for calibration, external validity, implementation feasibility, and alert management remains incomplete (Zubair *et al.* 2025; Zhang *et al.* 2026).

The clinical endpoint also matters. Many sepsis models predict mortality after sepsis is already recognized, which is useful for prognosis but different from anticipating the onset of Sepsis-3 physiology. Interpretable prognosis studies in MIMIC-IV have shown that tree-based models can identify clinically plausible drivers such as Glasgow Coma Scale, blood urea nitrogen, respiratory rate, urine output, and age (Hu *et al.* 2022). MIMIC-IV mortality prediction studies have also reported strong discrimination with reduced feature sets and ensemble models (Gao *et al.* 2024a). However, mortality prediction among already septic patients does not directly answer the bedside early-warning question: given the current patient state, will Sepsis-3 onset occur in a future time window?

Recent studies have therefore begun to focus on earlier and more operational targets. Liu et al. developed an interpretable MIMIC-IV emergency-triage model for sepsis risk and showed that comprehensive structured triage information outperformed vital signs alone, with Gradient Boosting reaching the highest AUC among the tested models (Liu *et al.* 2025). Duval et al. studied a related anticipatory task, predicting vasopressor initiation in ICU sepsis patients using an interpretable EHR-based design with matched case-control windows, and reported a Random Forest validation AUROC of 0.75 (Duval *et al.* 2025). These studies demonstrate the value of routinely collected physiology, but they also underline the importance of specifying the landmark time, the lookback window, and the event window with enough precision to avoid optimistic retrospective labeling.

Interpretability is not an optional cosmetic layer in this setting. A sepsis early-warning model may influence antibiotics, cultures, fluids, monitoring intensity, and ICU resource allocation; therefore, clinicians need to know whether the prediction is being driven by plausible physiology or by artifacts of documentation and ordering behavior. Recent MIMIC-IV interpretability work has shown that explanation methods can reveal model signals but can also expose demographic reliance and fairness concerns (Meng *et al.* 2022). Disease-specific interpretable studies have used SHAP to present global feature rankings, feature-effect plots, and individual explanations for sepsis prognosis and complications (Hu *et al.* 2022; Lu *et al.* 2022). In a sepsis forecasting study, global SHAP, local SHAP, and feature-dependence plots should therefore be reported alongside discrimination, calibration, and alarm-policy metrics.

A specific risk in sepsis forecasting is clinician-suspicion leakage. Variables such as newly ordered lactate, fresh coagulation tests, cultures, antibiotics, or vasopressor exposure may encode the bedside team's concern before the model observes overt organ dysfunction. If these signals are included without audit, the model may learn to reproduce clinical workflow rather than detect physiology earlier. The problem is particularly relevant for retrospective EHR studies because availability itself is informative: a missing or recently measured laboratory value can function as a proxy for clinical suspicion. Recent operational studies explicitly discuss realistic lookback windows, matching, and alert burden as ways to make retrospective prediction closer to real-time deployment (Duval *et al.* 2025; Boussina *et al.* 2024).

The present study is designed around that gap. We frame the prediction task as dynamic landmark forecasting: at each ICU time τ, the model estimates whether Sepsis-3 onset will occur in the interval $(\tau, \tau + h]$ for horizons $h \in \{4, 6, 12, 24\}$ hours. We use adult first ICU stays from MIMIC-IV v3.1, construct hourly physiologic features and temporal trend features, and train one LightGBM classifier per horizon. We then perform a paired feature-set audit by comparing a FULL model against a CLEAN model that removes the vasopressor flag, measurement-age channels, and observation-mask channels that can proxy clinician suspicion. Because deidentified admission-year group is not a bedside physiologic variable, the deployment-oriented primary analysis uses the no-admission-year CLEAN model, while the year-including CLEAN model is retained as an internal benchmark.

Our contribution is therefore not another single AUROC table, but a reproducible, physiology-focused sepsis forecasting pipeline. In the current MIMIC-IV v3.1 cohort, the no-year CLEAN model reaches held-out test AUROCs of 0.782, 0.772, 0.750, and 0.720 for 4-, 6-, 12-, and 24-hour horizons, respectively. At the 6-hour horizon, the recommended top-2 percent alarm policy with a 6-hour cooldown detects 17.8 percent of forecastable sepsis stays within the horizon-concordant 0–6 hour window, 29.3 percent within 0–24 hours, and 40.3 percent under a broader any-pre-onset definition. SHAP analysis identifies Glasgow Coma Scale total score, body temperature, short-window urine-output variability, and blood urea nitrogen as dominant predictors, aligning the model's behavior with organ dysfunction rather than measurement-process artifacts. This combination of leakage audit, multi-horizon forecasting, alarm-policy evaluation, and patient-level explanation is intended to support transparent analysis of Sepsis-3 early warning.

RELATED WORK

Systematic reviews and reporting gaps

Recent reviews show a rapidly expanding but methodologically heterogeneous sepsis prediction literature. Islam *et al.* (2023) reviewed EHR-based machine-learning and deep-learning studies for early sepsis prediction and found that most papers used longitudinal data, but that prediction windows, feature engineering, missing-data handling, and reporting quality varied substantially. Gao *et al.* (2024b) performed a systematic review and network meta-analysis of machine-learning algorithms in sepsis prediction and reported favorable rankings for deep and boosted approaches, while also emphasizing the need for stronger evaluation and algorithm comparison. Zubair *et al.* (2025) reviewed machine-learning and deep-learning approaches for sepsis diagnosis and highlighted the same translational barriers: EHR heterogeneity, class imbalance, interpretability, and limited real-world validation. Zhang *et al.* (2026) focused specifically on emergency department AI sepsis prediction and reported broad variation in AUROC, model type, feature count, and prediction windows across included studies. Our study follows these reviews by reporting not only AUROC but also AUPRC, Brier score, calibration, alarm burden, and lead time.

MIMIC-IV sepsis models and interpretability

Several closely related MIMIC-IV studies predict prognosis or complications after sepsis has been identified. Hu *et al.* (2022) developed interpretable machine-learning models for early in-hospital mortality prediction in critically ill sepsis patients and found that XGBoost performed best, with SHAP highlighting Glasgow Coma Scale, blood urea nitrogen, respiratory rate, urine output, and

age. Lu *et al.* (2022) developed an interpretable model for sepsis-associated encephalopathy, again using SHAP to connect model outputs to clinically relevant features. Gao *et al.* (2024a) reported a MIMIC-IV sepsis mortality model with a reduced feature set and high AUROC, emphasizing interpretability and feature reduction. Zhang *et al.* (2025) extended this prognosis theme to multiple outcomes, using MIMIC-IV to predict rapid recovery, chronic critical illness, and mortality, with CatBoost as the strongest model and urine output, respiratory rate, and temperature among the most important variables. These studies support the relevance of our top SHAP features, but their targets are prognosis among septic patients rather than pre-onset landmark forecasting.

Early warning, deployment, and positioning of the present study

Early-warning studies are closer to our target because they aim to identify sepsis before or near clinical recognition. Liu *et al.* (2025) used MIMIC-IV emergency triage data and compared vital-sign-only modeling with a broader structured EHR feature set; Gradient Boosting achieved the highest reported AUC and SHAP/LIME were used to explain the model. Boussina *et al.* (2024) evaluated the real-world deployment of COMPOSER and reported that deployment was associated with lower sepsis mortality and improved bundle compliance, making it one of the most important recent implementation studies in the field. Shashikumar *et al.* (2025) later integrated a large language model with COMPOSER to extract context from clinical notes for uncertain cases and reported improvements over the standalone model. Yang *et al.* (2026) proposed a semantic abstraction rule engine to transform structured clinical time series into clinician-readable text and fine-tuned lightweight LLMs for early sepsis prediction. These studies demonstrate growing interest in multimodal and language-model-based prediction, while our work keeps the primary model tabular and explainable to prioritize reproducibility, feature auditability, and fast SHAP computation.

The most direct MIMIC-IV onset comparison is the 2024 IEEE MedAI study by Shan *et al.* (2024), which predicted sepsis onset in ICU patients using multiple machine-learning models and 81 features extracted around ICU admission; XGBoost achieved an AUC of 0.81, and feature selection was used to simplify the model. That work is valuable as an onset-focused benchmark, but it uses an admission-centered feature window rather than repeated landmark forecasts over the ICU stay. Duval *et al.* (2025) addressed a neighboring anticipatory ICU problem by predicting vasopressor initiation several hours before therapy in Sepsis-3 patients, using matched windows and SHAP-based interpretation to reduce temporal and selection bias. Our study differs by directly forecasting Sepsis-3 onset at multiple future horizons, explicitly auditing clinician-suspicion proxy features, and evaluating practical alarm policies with cooldown and lead-time distributions.

Taken together, the recent literature supports four design choices for the present work. First, Sepsis-3 forecasting should be framed as a time-indexed prediction problem rather than a static admission classification task. Second, evaluation should include precision-recall behavior, calibration, lead time, and alert burden because sepsis prevalence at each landmark is low and bedside alerts have operational cost. Third, interpretability should include global and local views, not only a single feature-importance table. Fourth, EHR workflow proxies should be audited because they can inflate retrospective performance. The proposed CLEAN LightGBM pipeline combines these points: it uses MIMIC-IV v3.1, repeated landmarks, multi-horizon labels, a paired leakage-ablation experiment, SHAP explanations, and alarm-policy operating points, positioning the study as a reproducible clinical-ML benchmark rather than a black-box performance comparison.

METHODOLOGY

Study design

This study was designed as a retrospective, observational, dynamic prediction study using deidentified adult intensive care unit (ICU) data from MIMIC-IV v3.1 (Johnson *et al.* 2023, 2024). The objective was not to classify a patient once at admission, but to repeatedly estimate future Sepsis-3 onset risk at discrete ICU landmark times. At every eligible landmark time τ, the model receives only information available at or before τ and outputs the probability that Sepsis-3 onset will occur in a future horizon h. The evaluated horizons were $h \in \{4, 6, 12, 24\}$ hours (Figure 1).

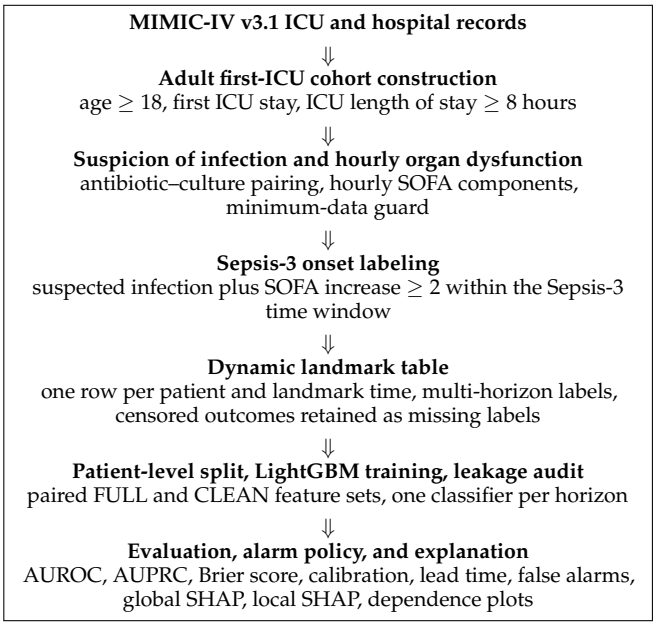


Figure 1 End-to-end methodological workflow. The workflow is intentionally organized so that outcome definitions are created before landmark sampling, patient-level splitting is performed before model fitting, and alarm/explanation analyses are applied only after held-out prediction

Detailed implementation tables, feature definitions, item identifiers, and software versions are provided in the supplementary reproducibility material. The main text retains only the tables needed to interpret cohort construction, model performance, operating points, and explanations.

Data source and cohort construction

The source population was drawn from MIMIC-IV v3.1, a deidentified critical-care database with linked hospital and ICU records (Table 1). The initial cohort was restricted to adult patients aged at least 18 years at ICU admission. Only the first ICU stay per patient was retained to avoid correlation from repeated ICU admissions. ICU stays shorter than 8 hours were excluded because they do not provide enough time for both early physiologic observation and future-event forecasting.

After the base ICU cohort was created, patients whose Sepsis-3 onset occurred within the first 4 hours of ICU admission were not used for forecasting. These patients are clinically important

but do not match the prospective prediction question, because the outcome is already present or nearly present at the first usable landmark. The remaining forecastable cohort contains ICU stays for which a future sepsis, death, or discharge trajectory can be evaluated from repeated landmarks.

Table 1 Cohort construction summary. Counts describe the final analytic pipeline used for the multi-horizon forecasting study

Cohort stage	Count	Interpretation
Adult first ICU stays with ICU length of stay ≥ 8 hours	64,363	Base analytic ICU cohort.
Suspicion-of-infection positive stays	32,521	Antibiotic–culture timing rule satisfied.
Sepsis-3 positive stays in the base cohort	29,004	Suspected infection plus acute organ dysfunction.
Forecastable stays after excluding very-early sepsis	42,291	Dynamic prediction cohort.
Forecastable Sepsis-3 positive stays	6,431	Future-onset sepsis cases available for prediction.
Landmark rows generated for model development	765,510	Repeated patient-time prediction examples.

Suspicion of infection

Suspicion of infection was operationalized as a valid temporal pairing between antimicrobial treatment and microbiologic culture, following the logic used in Sepsis-3 EHR studies (Singer *et al.* 2016; Seymour *et al.* 2016). Culture events were extracted from microbiology records. Antibiotic exposure was extracted from ICU medication administration and hospital pharmacy sources. To reduce under-detection, antimicrobial identification included both structured ICU antibiotic categories and medication-name matching using a broad antibiotic lexicon. This allowed intravenous ICU antibiotics, emergency department medications, oral medications, and other hospital medication records to contribute to the suspected infection signal when they satisfied the required time window.

For a culture-first pattern, the antibiotic had to occur within 24 hours after culture collection. For an antibiotic-first pattern, the culture had to occur within 72 hours after antibiotic initiation. The earliest valid pair was retained as the suspicion-of-infection time. Medication text fields were handled as medication strings rather than numeric-only quantities so that antimicrobial names could be matched reliably even when pharmacy quantity fields had heterogeneous data types.

Hourly SOFA construction and Sepsis-3 endpoint

Hourly Sequential Organ Failure Assessment (SOFA) components were constructed from routinely collected ICU physiology and laboratory data (Table 2). The respiratory component used oxygenation measures and mechanical-ventilation status so that higher respiratory SOFA grades were gated by ventilatory support when appropriate. The cardiovascular component used mean arterial pressure and dose-stratified vasoactive medication exposure. The renal component used serum creatinine and 24-hour urine output. The neurologic, coagulation, and liver components used Glasgow Coma Scale total score, platelet count, and total bilirubin, respectively.

Several harmonization steps were required before SOFA scoring. Fahrenheit temperature charting was converted to Celsius and merged with Celsius temperature charting. Laboratory lookback was allowed to extend to the hospital admission period so that early ICU SOFA did not discard clinically relevant pre-ICU laboratory information. Urine output was aggregated over clinically meaningful rolling windows. Mechanical ventilation and vasoactive medication exposure were mapped from structured ICU events. A minimum-data flag was then applied so that SOFA baselines and peaks were not determined from hours with too few observed organ-system components.

Table 2 Compact Sepsis-3 label algorithm used in the main analysis. The endpoint is an EHR-defined Sepsis-3 onset time, not an independently adjudicated biological state

Label component	Operational definition
Suspicion of infection	Earliest valid antibiotic–culture pair: culture followed by antibiotic within 24 hours, or antibiotic followed by culture within 72 hours. The suspicion time was the earlier timestamp in the valid pair.
SOFA window	Hourly SOFA rows inside $[T_{SI} - 48\text{h}, T_{SI} + 24\text{h}]$. ED and ward laboratory values from hospital admission through ICU time were allowed to support early SOFA reconstruction.
Baseline and acute increase	Baseline SOFA was the minimum hourly SOFA in the Sepsis-3 window; peak SOFA was the maximum. A stay met the organ-dysfunction criterion when peak minus baseline was at least 2 points.
Minimum-data guard	SOFA rows used for baseline/peak selection required at least 3 observed organ-system inputs among MAP, GCS, platelets, bilirubin, creatinine or urine output, and oxygenation.
Onset timestamp	If suspected infection and SOFA increase ≥ 2 were both satisfied, the EHR-defined Sepsis-3 onset time was set to the suspicion-of-infection time.
Predictor exclusion	Antibiotic administrations, cultures, and the suspicion-of-infection timestamp were used only for labels. They were not included as predictors in the landmark models.

Landmark prediction setup and target definition

The prediction problem was formulated using landmarking. For each forecastable ICU stay, prediction rows were emitted at repeated landmark times τ, beginning at ICU hour 4 and continuing every 2 hours until the earlier of ICU hour 72 or the last usable hour before ICU discharge or death. Each row represents the patient's state at a specific time and contains only predictors known at or before that time.

For each horizon $h \in \{4, 6, 12, 24\}$, the binary target was defined as:

$$y_h(\tau) = \begin{cases} 1, & \text{if } \tau < T_{\text{sepsis}} \leq \tau + h, \\ 0, & \text{if the patient is observed through } \tau + h \\ & \text{without Sepsis-3 onset,} \\ \text{missing}, & \text{if discharge or death occurs before } \tau + h \\ & \text{without observed Sepsis-3 onset.} \end{cases}$$

Rows with missing labels for a given horizon were retained in the landmark table but excluded from that horizon's supervised training and metric calculation. This prevents early discharge or death before the prediction window from being incorrectly counted as evidence that sepsis would not have occurred.

Feature engineering

Feature engineering was performed in chronological order to preserve temporal validity (Figure 2). First, current values were created using the most recent observed value available at or before each ICU hour. Second, measurement-age and observation-mask channels were created for the FULL feature set to record how recently each physiologic variable had been observed. Third, short-term trend features were attached at landmark time, including changes over 2, 6, and 12 hours and rolling ranges over 6 and 12 hours. These trend features were computed only from prior values and therefore do not use future information.

The feature vocabulary included demographics, admission metadata, vital signs, neurologic status, routine laboratory tests, renal output, treatment exposure, and temporal dynamics. Sex and hospital admission type were passed to LightGBM as native categorical variables rather than one-hot encoded variables. Deidentified admission-year group was retained only for the year-including internal benchmark; it was removed from the deployment-oriented primary CLEAN model because it may encode secular documentation, coding, ICU practice, case-mix, or COVID-era effects rather than bedside physiology. Missing numeric values were left as missing values in the primary LightGBM model because LightGBM learns default split directions for missingness during tree construction. Because this native handling can still encode subtle measurement-availability information, even after explicit observation masks and measurement-age channels are removed, a median-imputed sensitivity analysis was added in which numeric missing values were replaced using training-set medians and no missingness indicators were supplied.

Patient-level data splitting

The dynamic landmark table was split at the patient level into training, validation, and test partitions using a 70/15/15 ratio. Stratification was based on whether the patient ever developed forecastable Sepsis-3. Because each patient contributes multiple landmark rows, patient-level splitting prevents leakage in which earlier rows from a patient appear in training while later rows from the same patient appear in testing. The training, validation, and test partitions contained 531, 407, 117, 379, and 116, 724 landmark rows, respectively; the full split table is provided in the supplementary material.

LightGBM model development

One LightGBM gradient-boosted decision-tree classifier was trained for each prediction horizon (Ke *et al.* 2017). For a given horizon, only landmarks with observed labels were used for supervised fitting. The same training, validation, and test split assignment was reused across horizons. Hyperparameters were fixed across horizons to make the comparison interpretable: 1000 maximum boosting iterations, learning rate 0.03, 31 leaves, minimum child samples 100, feature fraction 0.8, bagging fraction 0.8, bagging frequency 5, and a fixed random seed. Early stopping used validation AUROC with a patience of 50 boosting rounds.

The 4-, 6-, 12-, and 24-hour models were treated as separate horizon-specific classifiers rather than a single multi-output network. This choice keeps each horizon's class balance, calibration, and alarm operating points explicit. The primary output of each model is a probability $p_h(\tau)$, interpreted as the estimated risk of Sepsis-3 onset within the next h hours conditional on the patient state at τ.

No row over-sampling, negative under-sampling, or focal-loss variant was used for the primary LightGBM models. The models were trained on the observed landmark distribution, and class imbalance was handled at evaluation and deployment stages through AUPRC reporting, validation-selected operating thresholds, and alarm-policy analysis. A regularized logistic-regression baseline used class-balanced weighting as a linear comparator, but this was not part of the primary model specification.

Leakage-ablation audit

Two paired feature sets were evaluated. The FULL feature set included all physiologic values, temporal features, measurement-age channels, observation-mask channels, and the active vasopressor flag. The CLEAN feature set removed 41 clinician-suspicion proxy channels: the vasopressor flag, 20 measurement-age channels, and 20 observation-mask channels. The same model specification was then trained on both feature sets for every horizon.

The leakage audit was defined as the performance difference between FULL and CLEAN models on the held-out test split. If removing workflow-proxy features caused only a small AUROC reduction, the model was interpreted as primarily dependent on physiologic state and trends rather than clinician-suspicion proxies. If the reduction was large, the model would be considered substantially dependent on workflow artifacts. The audit threshold was set before interpretation: a ΔAUROC below 0.02 across horizons would be considered compatible with a conservative CLEAN feature set.

Model evaluation

Discrimination was evaluated using AUROC. Because the positive class is rare at individual landmark times, AUPRC was also reported as a primary practical metric. Calibration and probabilistic accuracy were evaluated using Brier score and decile calibration. All metrics were calculated on the validation and held-out test splits, with the test split used for final reporting.

The full reporting plan includes cohort tables, label observability tables, per-horizon performance tables, calibration plots, receiver-operating characteristic curves, precision-recall curves, leakage-ablation tables, alarm operating-point tables, lead-time summaries, and SHAP explanation figures. This structure is intended to make the study auditable from raw cohort construction through patient-level alarm behavior.

Statistical uncertainty was quantified using a patient-level clustered bootstrap on the held-out test split. Each bootstrap replicate sampled ICU stays with replacement and included all eligible landmarks from the sampled stays, preserving within-patient row correlation. Percentile 95% confidence intervals were calculated

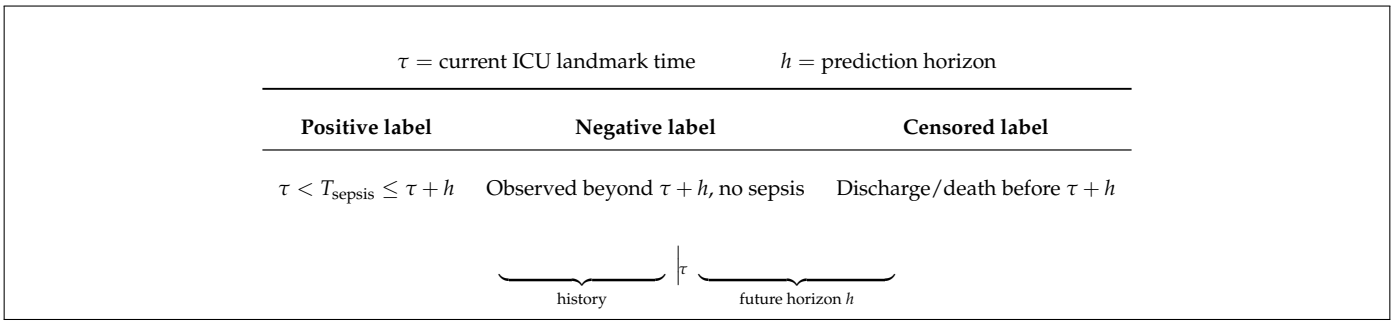


Figure 2 Landmark labeling scheme. The model sees only historical information up to τ, while the target asks whether Sepsis-3 onset occurs in the future interval $(\tau, \tau + h]$.

from 300 bootstrap replicates for AUROC, AUPRC, Brier score, calibration intercept, calibration slope, expected calibration error, and selected alarm-policy metrics.

Additional robustness analyses were defined after the initial internal validation to address concerns about temporal transportability, clinical-score baselines, missingness, lactate-related leakage, SOFA-component overlap, competing events, urine-output outliers, and subgroup performance. Temporal validation trained models on deidentified admission-year groups 2008–2016, selected early stopping on 2017–2019, and tested on 2020–2022 while excluding admission-year group from the feature set. Clinical score baselines included qSOFA, SIRS, current SOFA total, and change in SOFA from the early ICU period. Sensitivity models evaluated median imputation of numeric missing values, removal of lactate-derived features, removal of direct SOFA-component features, a composite sepsis-or-death endpoint, urine-feature winsorization, and subgroup analyses by sex, age group, admission type, deidentified admission-year group, first ICU care unit, and baseline severity.

Alarm-policy analysis

Raw probabilities were converted into alarm policies because a default probability threshold such as 0.5 is unsuitable for a low-prevalence sepsis forecasting problem. Thresholds were selected on the validation split and then applied unchanged to the test split. Candidate thresholds included top-risk row percentiles, validation-set sensitivity targets, and the threshold maximizing validation F1 score.

To mimic practical bedside alert suppression, cooldown windows of 0, 6, and 12 hours were evaluated. When cooldown was active, a patient could not emit another alert until the cooldown interval had elapsed. Alarm performance was summarized at both row level and patient level. Because patient-level first alerts can occur earlier than the nominal model horizon, detection was reported under four nested definitions: horizon-concordant detection, requiring an alert 0–6 hours before onset for the 6-hour model; early-warning detection within 0–24 hours; early-warning detection within 0–48 hours; and any pre-onset alert. Lead time was calculated as onset time minus alert time among detected sepsis stays. Non-sepsis false-alarm rate was defined as the proportion of non-sepsis stays with at least one emitted alarm. Alarm burden was reported as emitted alerts per 100 patient-days.

Explainability analysis

Explainability was performed with Tree SHAP, which computes additive feature attributions for tree-based models (Lundberg *et al.* 2020). Explanations were generated for the no-year CLEAN LightGBM models. For binary LightGBM, SHAP values were interpreted on the model log-odds scale; predicted risks were reported separately after transforming model output to probability. Global SHAP analysis summarized the mean absolute contribution of each feature across a held-out sample of 5,000 labeled landmarks. Because the same patient can contribute multiple landmarks, global SHAP was interpreted as landmark-level model behavior rather than as an independent patient-level causal ranking. Beeswarm plots were used to show attribution magnitude and direction for numeric features. Native categorical variables were interpreted by their SHAP contribution, although their beeswarm colors may appear neutral because categorical levels do not map to a continuous low-to-high color scale; categorical SHAP summaries by level are provided in the supplement.

Local SHAP analysis was performed for a representative held-out sepsis case at a landmark preceding Sepsis-3 onset. This local view decomposes the patient's predicted risk into positive and negative feature contributions at the chosen landmark. A temporal SHAP heatmap was also used for the same patient to show how explanations evolved across earlier landmarks. Finally, dependence plots were produced for the most influential numeric predictors to examine the learned feature-risk relationships.

RESULTS

Cohort yield and event structure

The base analytic cohort contained 64,363 adult first ICU stays with ICU length of stay at least 8 hours. The mean ICU length of stay was 3.57 days and the median ICU length of stay was 1.94 days. The mean age at ICU admission was 64.5 years and the median age was 66 years. Suspicion of infection was identified in 32,521 stays, corresponding to 50.5% of the base cohort. Sepsis-3 criteria were met in 29,004 stays, corresponding to 45.1% of the base cohort and 89.2% of suspicion-of-infection positive stays. After excluding sepsis present at or near ICU admission, 42,291 stays formed the forecastable cohort, of which 6,431 were forecastable Sepsis-3 positive stays. After landmark feasibility filtering, 41,166 stays contributed 765,510 landmark rows to model development. Detailed baseline characteristics by final event type are provided in the supplementary material.

Among landmark-contributing stays, the final event type distribution was 5,374 sepsis stays, 1,645 death-without-forecastable-sepsis stays, and 34,147 discharge-without-forecastable-sepsis stays. Landmark rows were highly imbalanced at every horizon. The observed positive prevalence increased with longer horizon length, from 0.87% at the 4-hour test horizon to 5.87% at the 24-hour test horizon, while censoring also increased with horizon

length because longer windows more often extended beyond discharge or death. The full label-observability table by split and horizon is included in the supplementary material.

Multi-horizon model discrimination and calibration

The deployment-oriented no-year CLEAN LightGBM models showed consistent temporal degradation as the prediction horizon lengthened (Table 3). On the held-out test split, AUROC was 0.782 for 4-hour prediction, 0.772 for 6-hour prediction, 0.750 for 12-hour prediction, and 0.720 for 24-hour prediction. AUPRC increased with horizon length because event prevalence also increased with the longer prediction windows: 0.050 at 4 hours, 0.064 at 6 hours, 0.091 at 12 hours, and 0.145 at 24 hours. Brier scores remained numerically low, reflecting the rare-event probability scale, but increased from 0.0085 at 4 hours to 0.0530 at 24 hours.

Table 3 Deployment-oriented no-year CLEAN LightGBM performance by prediction horizon on the held-out test split. N denotes rows with observed labels for the corresponding horizon

Horizon	Split	N	Positive	Prevalence	AUROC	AUPRC	Brier
4h	Test	110,365	962	0.87%	0.782	0.050	0.0085
6h	Test	106,040	1,450	1.37%	0.772	0.064	0.0131
12h	Test	93,293	2,595	2.78%	0.750	0.091	0.0262
24h	Test	71,818	4,218	5.87%	0.720	0.145	0.0530

Patient-level clustered bootstrap confidence intervals supported the stability of the main discrimination estimates while appropriately accounting for repeated landmarks within patients. At the primary 6-hour horizon, AUROC was 0.772 (95% CI 0.752–0.792), AUPRC was 0.064 (95% CI 0.055–0.077), Brier score was 0.0131 (95% CI 0.0123–0.0142), and calibration slope was 1.10 (95% CI 1.02–1.16). The full bootstrap table is included in the supplementary material (Figure 3).

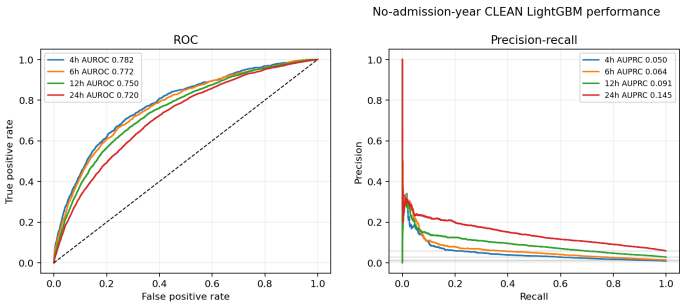


Figure 3 Held-out ROC and precision–recall curves for the no-year CLEAN models across all prediction horizons. Precision–recall panels include prevalence baselines, which are essential for interpreting rare-event prediction. Longer horizons have higher event prevalence but lower AUROC, consistent with the increasing uncertainty of farther-ahead onset prediction

At the 6-hour horizon, decile calibration showed a monotonic increase in observed event frequency as predicted risk increased (Table 4). The lowest decile had mean predicted risk 0.0035 and observed frequency 0.0015, while the highest decile had mean predicted risk 0.0500 and observed frequency 0.0560 (Figure 4). Thus, the model concentrated most sepsis events in the upper risk tail while preserving an interpretable probability scale. The full decile-calibration table is provided in the supplementary material.

Table 4 Held-out test performance with patient-level clustered bootstrap 95% confidence intervals.

Metric	4 h	6 h	12 h	24 h
Prevalence	0.87%	1.37%	2.78%	5.87%
AUROC	0.782 (0.763–0.801)	0.772 (0.752–0.792)	0.750 (0.729–0.769)	0.720 (0.699–0.742)
AUPRC	0.050 (0.042–0.061)	0.064 (0.055–0.077)	0.091 (0.079–0.105)	0.145 (0.128–0.165)
Brier	0.0085 (0.0078–0.0090)	0.0131 (0.0123–0.0142)	0.0262 (0.0241–0.0282)	0.0530 (0.0485–0.0577)
Slope	1.00 (0.94–1.06)	1.10 (1.02–1.16)	1.07 (0.98–1.16)	1.06 (0.95–1.17)
Intercept	−0.01 (−0.19, 0.25)	0.37 (0.06, 0.60)	0.21 (−0.08, 0.50)	0.09 (−0.17, 0.34)
ECE10	0.0007 (0.0006–0.0015)	0.0020 (0.0012–0.0030)	0.0028 (0.0018–0.0048)	0.0037 (0.0029–0.0089)

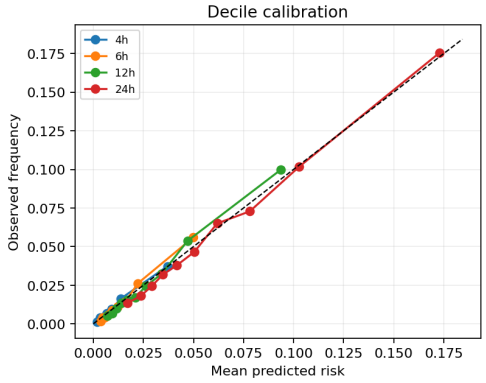


Figure 4 Decile calibration for the no-year CLEAN models across all prediction horizons. Points compare mean predicted risk with observed event frequency in each decile

Baseline, temporal-validation, and sensitivity analyses

The no-year CLEAN LightGBM model outperformed reconstructed clinical scores and a regularized logistic-regression comparator at the primary 6-hour horizon. qSOFA, SIRS, current SOFA total, and early SOFA change had AUROCs between 0.567 and 0.628 (Table 5). A class-weighted regularized logistic model using the CLEAN feature set reached AUROC 0.699 and AUPRC 0.031; because its raw probabilities were poorly calibrated, Platt calibration was fitted on the validation split before reporting Brier score. In contrast, the no-year CLEAN LightGBM model reached AUROC 0.772 and AUPRC 0.064 at the same horizon. This supports the added value of nonlinear tabular modeling and temporal feature interactions beyond simple score thresholds or a linear model.

Table 5 Primary 6-hour model compared with clinical-score and machine-learning baselines on the held-out test split. Brier score is not reported for raw ordinal scores because they are not calibrated probabilities

Comparator	AUROC	AUPRC	Brier	Notes
qSOFA	0.608	0.019	N/A	Ordinal clinical score
SIRS	0.567	0.017	N/A	Ordinal clinical score
Current SOFA total	0.614	0.019	N/A	EHR-reconstructed SOFA at landmark
SOFA change from early ICU period	0.628	0.022	N/A	Early-delta clinical score
Regularized logistic regression	0.699	0.031	0.0134	Platt-calibrated linear CLEAN-feature model
No-year CLEAN LightGBM	0.772	0.064	0.0131	Deployment-oriented nonlinear model

Temporal validation produced slightly lower, but still coherent, discrimination when models were trained on earlier deidentified admission-year groups, validated on 2017–2019, and tested on 2020–2022 after removing admission-year group from the predictors (Table 6). At the 6-hour horizon, temporal-test AUROC

Table 6 Temporal validation using earlier deidentified admission-year groups for training, 2017–2019 for validation, and 2020–2022 for testing. Admission-year group was excluded from temporal-validation predictors. Confidence intervals are patient-level clustered bootstrap intervals.

Metric	4 h	6 h	12 h	24 h
Temporal test *N*	133,859	129,579	116,908	94,774
Positive	1,246	1,897	3,598	6,177
Prevalence	0.93%	1.46%	3.08%	6.52%
AUROC (95% CI)	0.776 (0.759–0.791)	0.762 (0.746–0.777)	0.739 (0.722–0.756)	0.697 (0.679–0.715)
AUPRC (95% CI)	0.034 (0.030–0.040)	0.047 (0.042–0.052)	0.084 (0.076–0.094)	0.139 (0.126–0.152)
Brier (95% CI)	0.0093 (0.0087–0.0099)	0.0144 (0.0135–0.0153)	0.0292 (0.0272–0.0314)	0.0591 (0.0549–0.0628)

was 0.762 (95% CI 0.746–0.777), AUPRC was 0.047 (95% CI 0.042–0.052), and Brier score was 0.0144 (95% CI 0.0135–0.0153). The 24-hour temporal AUROC was 0.697 (95% CI 0.679–0.715), indicating greater degradation for farther-ahead prediction under temporal shift. These results are weaker than random patient-level testing, especially for AUPRC, but support the manuscript's positioning as an internally and temporally validated single-database benchmark rather than an externally validated deployment model.

Temporal alarm-policy behavior was also evaluated for the 6-hour no-year model. With a top-2% validation threshold and 6-hour cooldown, temporal-test horizon-concordant detection was 15.1%, 0–24 hour detection was 28.9%, and any-pre-onset detection was 46.3%. The temporal false-alarm rate was higher than in the random test split at 21.4%, with 17.6 alerts per 100 patient-days. Patient PPV was 9.2% for horizon-concordant detection and 17.7% for 0–24 hour detection, indicating that operating thresholds would need site- and period-specific monitoring before deployment. The full temporal alarm table is provided in the supplementary material.

Sensitivity analyses addressed residual reviewer concerns about missingness, lactate, SOFA-component overlap, deidentified admission-year metadata, competing events, and urine-output outliers (Table 7). At the 6-hour horizon, the year-including internal benchmark reached AUROC 0.779, while the no-year primary model reached AUROC 0.772. Median-imputing all numeric missing values in the CLEAN model reduced AUROC to 0.766, suggesting that native LightGBM missingness handling contributes modestly but does not dominate performance. Removing lactate-derived features produced AUROC 0.776, indicating that the model does not rely strongly on lactate availability or lactate values. Removing direct SOFA-component features produced a larger but still moderate reduction to AUROC 0.758, consistent with the model using early organ-dysfunction patterns that precede formal Sepsis-3 threshold crossing. A no-year urine-winsorized model, which clipped urine-related features at the training-set 1st–99th percentiles, reached AUROC 0.774 (Table 8). A no-year composite endpoint model for Sepsis-3 onset or death within 6 hours reached AUROC 0.808 and AUPRC 0.140, showing that the framework also captures broader deterioration risk but that this is a different clinical endpoint.

Subgroup analysis at the 6-hour horizon showed broadly similar discrimination across sex, age, and admission-year groups, but it also identified heterogeneity that would require prospective monitoring before deployment. AUROC was 0.770 in female patients and 0.786 in male patients. Age-stratified AUROC declined from 0.824 in patients aged 18–49 years to 0.751 in patients aged 75 years or older. Deidentified admission-year group performance varied from 0.725 in 2014–2016 to 0.821 in 2020–2022. The full subgroup table, including admission type, ICU type, baseline-severity group, calibration, and AUPRC, is provided in the supplementary material.

Table 7 Reviewer-directed 6-hour sensitivity analyses on the held-out test split. The sepsis-or-death row uses a retrained composite endpoint model and should not be compared as the same target. Full per-horizon sensitivity results are included in the supplementary material

6-hour feature sensitivity	Features	AUROC	AUPRC	Brier
No-year CLEAN LightGBM	129	0.772	0.064	0.0131
Year-including internal benchmark	130	0.779	0.073	0.0131
Median-imputed numeric missing values	130	0.766	0.066	0.0131
No lactate-derived features	124	0.776	0.074	0.0131
No direct SOFA-component features	88	0.758	0.069	0.0131
No-year urine winsorization	129	0.774	0.064	0.0132
No-year sepsis-or-death endpoint	129	0.808	0.140	0.0161

Table 8 Leakage-ablation discrimination results on the held-out test split. Negative deltas indicate lower CLEAN performance relative to FULL. Confidence intervals for ΔAUROC are patient-level paired bootstrap intervals. Calibration metrics are reported separately in Table 9.

Metric	4 h	6 h	12 h	24 h
AUROC (FULL)	0.793	0.784	0.757	0.731
AUROC (CLEAN)	0.785	0.779	0.755	0.728
ΔAUROC (95% CI)	−0.008 (−0.014, −0.002)	−0.005 (−0.010, +0.001)	−0.003 (−0.007, +0.002)	−0.002 (−0.007, +0.002)
AUPRC (FULL)	0.066	0.077	0.104	0.163
AUPRC (CLEAN)	0.057	0.073	0.099	0.154
ΔAUPRC	−0.009	−0.004	−0.005	−0.009

Subgroup alarm summaries showed similar heterogeneity in operating behavior. For the top-2% 6-hour policy, 0–24 hour detection was 22.7% in female patients and 30.2% in male patients, while non-sepsis false-alarm rates were 13.6% and 15.5%, respectively. Across age groups, 0–24 hour detection was highest in patients aged 18–49 years (30.7%) and lowest in patients aged 75 years or older (23.8%); the oldest group also had the highest non-sepsis false-alarm rate (17.7%). Across deidentified admission-year groups, 0–24 hour detection ranged from 22.3% in 2014–2016 to 29.8% in 2011–2013, and alerts per 100 patient-days ranged from 10.6 to 18.9. These differences do not invalidate the aggregate model, but they indicate that subgroup calibration and alert burden would need active monitoring before prospective use.

Leakage-ablation results

The leakage-ablation analysis compared the FULL feature set against the CLEAN feature set after removing clinician-suspicion and treatment-proxy channels (Table 9). Across all horizons, the AUROC reduction after removing these channels was small: -0.008 at 4 hours, -0.005 at 6 hours, -0.003 at 12 hours, and -0.002 at 24 hours. Patient-level paired bootstrap intervals for the AUROC deltas were narrow and all centered near zero. The largest AUPRC reduction was -0.009. These changes are below the 0.02 AUROC concern threshold, supporting the interpretation that the model's performance is largely driven by physiology and prior physiologic trends rather than by measurement-process proxies. This leakage audit isolates workflow-proxy removal; the no-year primary model then additionally removes deidentified admission-year metadata for deployment-oriented reporting.

Table 9 Leakage-ablation calibration results on the held-out test split. All Brier-score increases after CLEAN feature filtering are below 3×10^{-4}, so probability calibration is essentially preserved

Horizon	Brier FULL	Brier CLEAN	ΔBrier
4h	0.00839	0.00844	+0.00004
6h	0.01306	0.01309	+0.00004
12h	0.02603	0.02610	+0.00007
24h	0.05250	0.05274	+0.00024

Alarm-policy results

The 6-hour model was selected as the primary operating horizon because it provides a clinically interpretable risk-enrichment window while preserving adequate positive-event density. At the 6-hour horizon, the no-year top-2% alarm policy with 6-hour cooldown detected 17.8% of sepsis cases within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition. The same policy produced a 16.5% non-sepsis false-alarm rate and 15.50 alerts per 100 patient-days (Figure 5). Because any-pre-onset detection can include alerts earlier than the nominal 6-hour horizon, the horizon-concordant and early-warning definitions are the primary clinical interpretation.

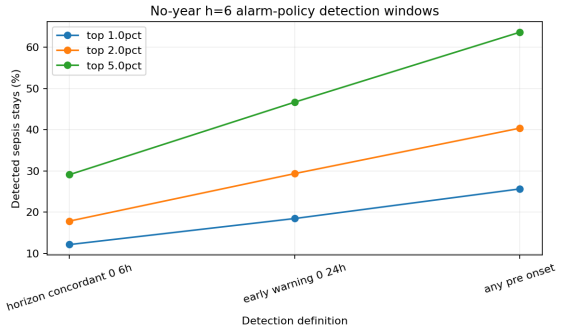


Figure 5 Alarm-policy detection for the 6-hour no-year CLEAN model under alternative lead-time definitions. The same validation-derived threshold and 6-hour cooldown can be interpreted as horizon-concordant detection (0–6h before onset), broader early-warning detection (0–24h or 0–48h), or any pre-onset enrichment

The revised detection-window analysis clarifies the relationship between a 6-hour prediction model and patient-level alert timing (Table 10). For the top-2% no-year policy, horizon-concordant detection within 0–6 hours before onset was 17.8%, with a median lead time of 2.34 hours. The same policy detected 29.3% of sepsis stays within a broader 0–24 hour early-warning window and 32.8% within 0–48 hours. The 40.3% detection corresponds to the broad any-pre-onset definition, which is useful for risk enrichment but should not be interpreted as strictly horizon-concordant 6-hour detection. The denominator distinction is important: the h=6 test split contained 106,040 observed landmark rows and 1,450 positive landmark rows, but patient-level alarm detection used 808 sepsis-positive ICU stays as the denominator. The top-2% policy alerted 1,221 ICU stays and emitted 1,508 cooldown-filtered alerts; this corresponds to 10.5 alerts per horizon-concordant detected case and Patient PPV of 11.8% under the strict 0–6 hour definition.

Table 10 No-year h=6 alarm-policy analysis under horizon-concordant, early-warning, and broad any-pre-onset definitions. Patient PPV is calculated as detected sepsis stays divided by all alerted stays for that definition

Policy	Detection definition	Caught / 808	Detection	Median lead	Patient PPV	False alarm	Alerts / 100 pt-d
Top 1%	Horizon-concordant 0–6h	98	12.1%	2.35 h	13.8%	9.3%	8.21
Top 1%	Early-warning 0–24h	149	18.4%	4.00 h	20.9%	9.3%	8.21
Top 1%	Any pre-onset	207	25.6%	7.63 h	29.1%	9.3%	8.21
Top 2%	Horizon-concordant 0–6h	144	17.8%	2.34 h	11.8%	16.5%	15.50
Top 2%	Early-warning 0–24h	237	29.3%	4.87 h	19.4%	16.5%	15.50
Top 2%	Any pre-onset	326	40.3%	9.44 h	26.7%	16.5%	15.50
Top 5%	Horizon-concordant 0–6h	235	29.1%	2.58 h	10.7%	31.1%	34.16
Top 5%	Early-warning 0–24h	377	46.7%	5.83 h	17.1%	31.1%	34.16
Top 5%	Any pre-onset	514	63.6%	11.41 h	23.4%	31.1%	34.16

Global SHAP analysis

Global SHAP analysis of the 6-hour no-year CLEAN model identified neurologic status, temperature, admission route, renal-output dynamics, renal laboratory values, inflammatory markers, acid-base dynamics, and vital signs as the strongest contributors (Table 11). Glasgow Coma Scale total score was the dominant predictor by mean absolute SHAP value. Body temperature, hospital admission type, range of 6-hour urine output over the previous 12 hours, blood urea nitrogen, white blood cell count, bicarbonate change, respiratory rate, and heart rate followed. Categorical variables such as hospital admission type are shown with neutral coloring in beeswarm visualizations because they are native categorical LightGBM inputs rather than ordered continuous variables; a separate categorical SHAP summary by level is included in the supplementary material.

SHAP rankings were interpreted as model-attribution summaries rather than causal rankings (Figure 7). Several predictors are physiologically correlated, including blood urea nitrogen, creatinine, urine output, and SOFA-related renal dysfunction; similarly, heart rate, respiratory rate, temperature, and white blood cell count can co-vary during systemic inflammation. Correlation can distribute or shift SHAP credit among related variables, so global rankings were interpreted together with dependence plots, local explanations, and clinical domain knowledge.

Local SHAP example

A representative held-out sepsis case was selected for local explanation (Table 12). At ICU hour $\tau = 8$, the 6-hour no-year model assigned a predicted Sepsis-3 onset risk of 0.452 (Figure 7). The patient's Sepsis-3 onset occurred 4.35 hours after the landmark, placing the event inside the 6-hour prediction window. The largest positive contribution was Glasgow Coma Scale total score of 3, followed by hemoglobin 14.7 g/dL, body temperature 37.56 °C, lactate 3.5 mmol/L, 6-hour lactate range, emergency admission type, blood glucose 174 mg/dL, total bilirubin 0.6 mg/dL, white blood cell count 20.2 K/uL, heart-rate range, urine-output variability, respiratory rate, and heart rate. The explanation is compatible with the patient's bedside pattern: depressed neurologic status, elevated lactate, leukocytosis, borderline hypotension, tachycardia, and stress hyperglycemia all push the prediction toward imminent Sepsis-3 onset. The locally large hemoglobin contribution should be interpreted as context-dependent model attribution, not as an isolated causal claim; correlated severity, fluid status, sampling context, and interactions with other features can shift SHAP credit for a single patient (Figure 8).

Table 11 Top 20 global SHAP features for the 6-hour no-year CLEAN model. Values are mean absolute SHAP contributions on the LightGBM log-odds scale across a held-out landmark sample

Rank	Feature	Mean |SHAP|
1	Glasgow Coma Scale total score	0.309
2	Body temperature	0.123
3	Hospital admission type	0.089
4	Range of 6-hour urine output over previous 12 hours	0.089
5	Blood urea nitrogen	0.076
6	White blood cell count	0.063
7	Change in serum bicarbonate over previous 12 hours	0.057
8	Respiratory rate	0.049
9	Heart rate	0.048
10	Change in hemoglobin over previous 12 hours	0.046
11	Change in white blood cell count over previous 6 hours	0.038
12	Change in serum creatinine over previous 12 hours	0.033
13	Hemoglobin	0.033
14	Total bilirubin	0.032
15	Change in white blood cell count over previous 12 hours	0.030
16	Change in peripheral oxygen saturation over previous 6 hours	0.030
17	Range of lactate over previous 6 hours	0.029
18	Age at ICU admission	0.029
19	Range of lactate over previous 12 hours	0.027
20	Serum sodium	0.025

Table 12 Top local SHAP contributors for a held-out sepsis patient at ICU hour 8. The no-year model predicted 45.2% risk of Sepsis-3 onset within 6 hours; onset occurred 4.35 hours later. SHAP values are log-odds contributions

Rank	Feature	Value at $\tau = 8$h	SHAP	|SHAP|
1	Glasgow Coma Scale total score	3.0	1.016	1.016
2	Hemoglobin	14.7 g/dL	0.336	0.336
3	Body temperature	37.56 °C	0.278	0.278
4	Lactate	3.5 mmol/L	0.276	0.276
5	Range of lactate over previous 6 hours	3.5 mmol/L	0.223	0.223
6	Hospital admission type	Emergency ward admission	0.152	0.152
7	Blood glucose	174 mg/dL	0.149	0.149
8	Total bilirubin	0.6 mg/dL	0.145	0.145
9	White blood cell count	20.2 K/uL	0.133	0.133
10	Range of heart rate over previous 12 hours	3 beats/min	0.127	0.127
11	Range of 6-hour urine output over previous 12 hours	150 mL	0.117	0.117
12	Respiratory rate	22 breaths/min	0.116	0.116
13	Heart rate	105 beats/min	0.099	0.099
14	Range of respiratory rate over previous 12 hours	1 breath/min	0.096	0.096
15	Urine output in previous 6 hours	116 mL	0.092	0.092

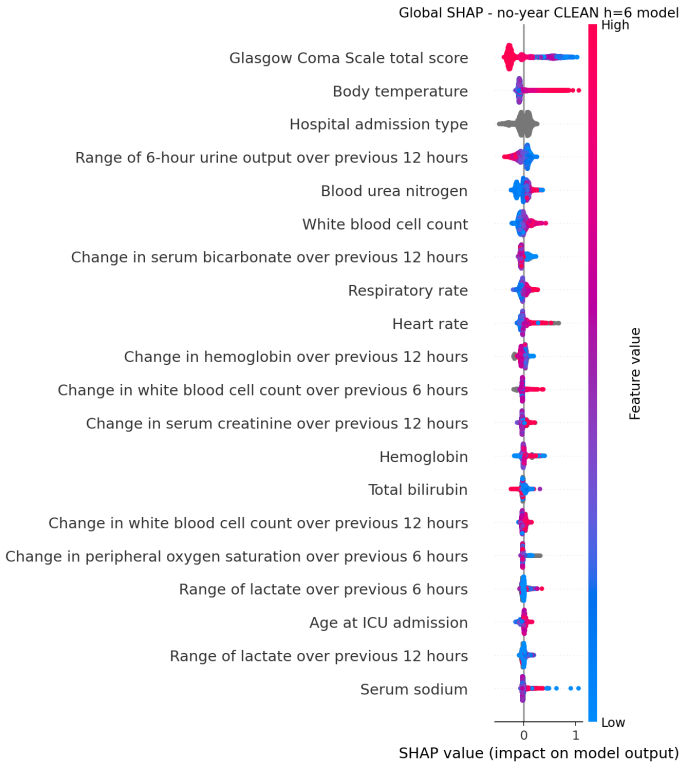


Figure 6 Global SHAP beeswarm plot for the primary 6-hour no-year CLEAN model. Other horizon-specific beeswarm plots and categorical SHAP summaries are provided in the supplementary figure set. SHAP values are on the model log-odds scale

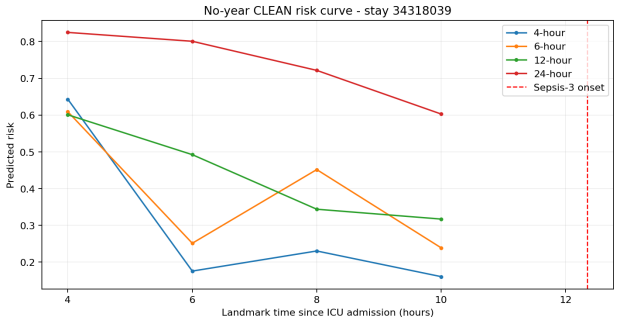


Figure 7 Patient-level 6-hour risk curve for the representative held-out sepsis case. The selected landmark is ICU hour 8, and Sepsis-3 onset occurred 4.35 hours later.

Dependence analysis of the four most influential numeric variables

Dependence plots were generated for the four most influential numeric predictors in the 6-hour no-year CLEAN model: Glasgow Coma Scale total score, body temperature, range of 6-hour urine output over the previous 12 hours, and blood urea nitrogen (Table 13). Lower Glasgow Coma Scale scores strongly increased risk, while a normal score of 15 was protective. Higher temperature values, especially above approximately 37.4 °C in the held-out distribution, increased risk. Lower 12-hour urine-output variability increased risk, consistent with persistently low or nonrecovering urine output, whereas very high variability was associated with lower SHAP values. Blood urea nitrogen showed a broadly

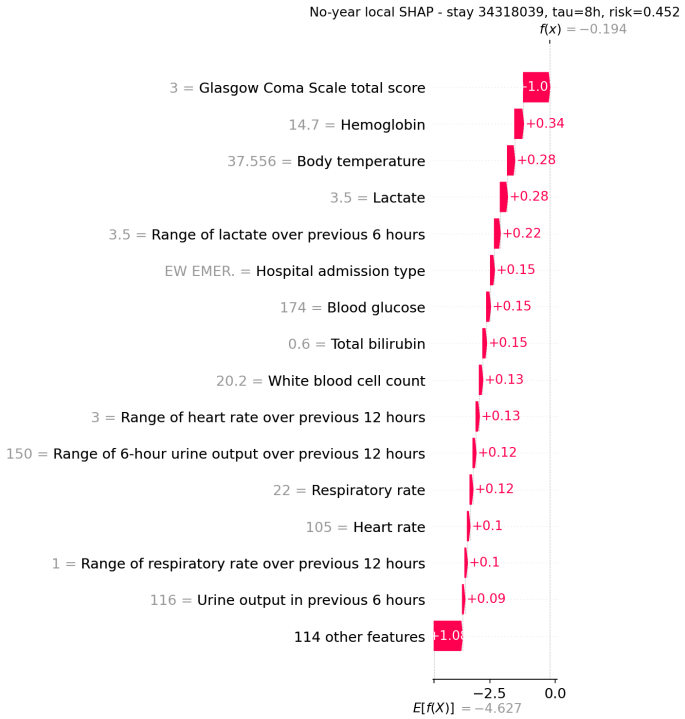


Figure 8 Local SHAP waterfall for the representative held-out sepsis case at ICU hour 8. The waterfall decomposes the selected landmark prediction on the LightGBM log-odds scale; predicted probability is reported separately in the text and table. The temporal SHAP heatmap is provided in the supplementary figure set

monotonic risk increase, with high values carrying positive SHAP contributions.

Table 13 Dependence-plot summary for the four most influential numeric SHAP features in the 6-hour CLEAN model. Ranges correspond to empirical bins from held-out landmarks. Per-feature clinical interpretations are given in the surrounding paragraph; the full quantitative summary with all 10 deciles per feature is provided in the supplementary material

Feature	Low / reference range	Mean SHAP	High-risk range	Mean SHAP
Glasgow Coma Scale total score	3–7 points	+0.649	15 points	−0.262
Body temperature	36.5–36.6 °C	−0.090	37.44–39.89 °C	+0.512
6 h urine output range over prior 12 h	0–99 mL	+0.112	1,560–4,258 mL	−0.255
Blood urea nitrogen	1–7 mg/dL	−0.098	48–181 mg/dL	+0.149

Summary of principal findings

The results support four main conclusions (Figure 9). First, Sepsis-3 onset forecasting is feasible from routinely collected ICU physiology, but the task remains difficult because landmark-level prevalence is low and farther horizons are less discriminable. Second, the CLEAN model retains almost all FULL-model discrimination after removal of measurement-process and treatment-proxy channels, reducing concern that performance is driven by clinician-suspicion leakage. Third, the 6-hour no-year model provides a risk-enrichment operating horizon only when detection windows

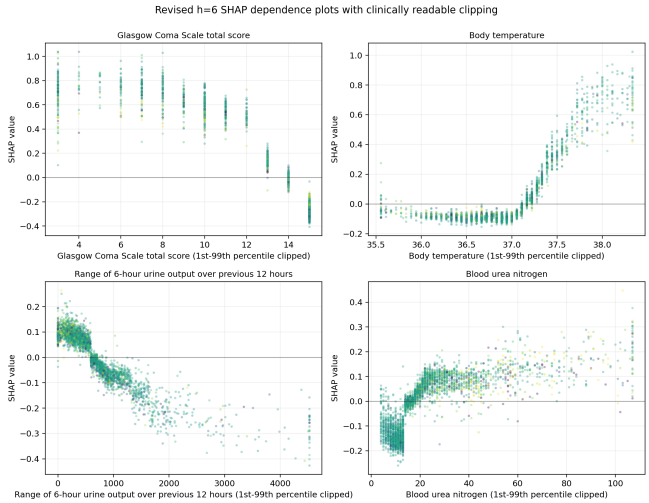


Figure 9 Revised SHAP dependence plots for the four most influential numeric predictors at the 6-hour horizon. Axes include clinical units and are clipped to the 1st–99th percentile for readability, reducing distortion from extreme urine-output values while preserving the learned nonlinear relationship between feature value and log-odds contribution to predicted Sepsis-3 risk

are stated explicitly: the top-2% alarm policy detects 17.8% of test sepsis stays within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition. Fourth, SHAP explanations are interpretable at both global and local levels, with dominant contributions from neurologic status, temperature, renal-output dynamics, blood urea nitrogen, inflammatory markers, vital signs, and lactate-related dynamics.

DISCUSSION

Principal interpretation

This study shows that Sepsis-3 onset can be forecast from routinely collected ICU physiology using a leakage-audited, multi-horizon tabular model, but it also shows why this task should not be interpreted like a static diagnosis or mortality classification problem. The deployment-oriented no-year CLEAN LightGBM model achieved held-out AUROCs of 0.782, 0.772, 0.750, and 0.720 for 4-, 6-, 12-, and 24-hour horizons, respectively. This pattern is clinically and statistically expected: as the horizon lengthens, the model is asked to predict a more remote and more uncertain physiologic transition. The fact that AUROC decreases with longer horizons while AUPRC increases reflects two simultaneous forces: farther-ahead prediction is harder, but longer windows contain more positive events and therefore have higher observed prevalence.

The main methodological finding is that removing clinician-suspicion proxy features produced only minimal performance loss. The largest AUROC difference between FULL and CLEAN models was -0.008 at the 4-hour horizon, and the 6-hour model lost only -0.005 AUROC. This is important because retrospective EHR sepsis prediction is vulnerable to inflated performance when models exploit signals such as recently ordered lactate, recent coagulation tests, culture ordering, or vasopressor exposure. In this pipeline, measurement-age channels, observation-mask channels, and the active vasopressor flag were removed from the primary model, yet performance remained stable. The resulting interpretation is that

the model is learning from physiologic state and prior physiologic dynamics rather than simply reproducing bedside workflow.

Comparison with model-performance literature

The performance level of the present model is best understood in relation to the exact prediction target. Studies that predict mortality after sepsis has already been identified often report higher discrimination than our onset-forecasting model. For example, Hu *et al.* (2022) reported an XGBoost AUC of 0.884 for early in-hospital mortality prediction among ICU patients with sepsis, and Gao *et al.* (2024a) reported a very high AUROC for sepsis mortality prediction using a reduced MIMIC-IV feature set. These tasks are clinically valuable, but they are not equivalent to predicting future Sepsis-3 onset in a mixed ICU population. Mortality models can use information from an already septic cohort, often after substantial organ dysfunction has appeared. Our target is earlier and stricter: at each landmark, the patient may not yet satisfy Sepsis-3 criteria, and the model must estimate whether onset will occur in a future window.

Compared with studies that are closer to onset prediction, our results are in a plausible range. Liu *et al.* (2025) developed an interpretable MIMIC-IV emergency-triage sepsis model and reported that Gradient Boosting achieved an AUC of 0.83 when comprehensive triage information was used. Shan *et al.* (2024) predicted sepsis onset in ICU patients using MIMIC-IV and reported an XGBoost AUC of 0.81 using admission-centered features. Our 4- and 6-hour AUROCs are slightly lower than these values, but our design differs in ways that make the task more conservative: repeated landmark prediction across the ICU stay, censoring-aware horizon labels, patient-level splitting, removal of workflow-proxy channels, and exclusion of deidentified admission-year group from the primary model. Duval *et al.* (2025) predicted vasopressor initiation in Sepsis-3 patients several hours before therapy and reported a Random Forest validation AUROC of 0.75; our 6-hour AUROC of 0.772 is similar in magnitude but targets Sepsis-3 onset rather than hemodynamic intervention.

The supplementary material includes a structured comparison table summarizing these related studies and the limits of direct metric comparison.

These comparisons reinforce a key point emphasized by recent systematic reviews: sepsis prediction studies are difficult to compare unless the target, prediction window, data source, censoring rule, and implementation setting are aligned (Islam *et al.* 2023; Gao *et al.* 2024b; Zubair *et al.* 2025; Zhang *et al.* 2026). A model with a higher AUROC may be solving a less prospective task, using a richer or less controlled feature set, or predicting a more downstream endpoint. For that reason, the most meaningful contribution of the present work is not that it outperforms every previous model numerically; rather, it offers a transparent, leakage-audited benchmark for dynamic Sepsis-3 onset forecasting.

The added baseline analyses further clarify the model's contribution. At the 6-hour horizon, qSOFA, SIRS, current SOFA, and early SOFA change achieved AUROCs between 0.567 and 0.628, while regularized logistic regression using the same CLEAN feature set reached AUROC 0.699. The no-year CLEAN LightGBM model reached AUROC 0.772 and AUPRC 0.064. This does not establish algorithmic novelty, but it does show that nonlinear tree-based modeling and engineered temporal features add measurable value over simpler clinical scores and a linear comparator under the same landmark target.

Interpretation of the horizon-specific results

The 4-hour model produced the highest AUROC, which is expected because near-term onset is closer to the current physiologic state. However, the 4-hour label was extremely rare at the landmark level, with only 0.87% observed test prevalence. This scarcity limits AUPRC and makes fixed high-probability thresholds unrealistic. The 24-hour model had lower AUROC but higher AUPRC because positive labels were more frequent across the longer window. The 6-hour horizon provided the most balanced operating point: it retained near-term discrimination while giving enough time for clinical reassessment, cultures, antibiotics, fluids, closer monitoring, or escalation.

The practical alarm results support this choice, but only when the detection window is stated explicitly. At the h=6 top-2% operating point with a 6-hour cooldown, the no-year model detected 17.8% of sepsis stays within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under the broad any-pre-onset definition, with 15.5 alerts per 100 patient-days. This is better interpreted as patient risk enrichment than as high-sensitivity bedside detection. The top-5% policy detected more patients but nearly doubled the false-alarm rate and alert burden. The top-1% policy had lower burden but missed more cases. This is consistent with the implementation literature: early-warning models can only help if alert rates are acceptable and the receiving clinical team trusts the signal (Boussina *et al.* 2024).

The revised alarm-window analysis also changes the interpretation of lead time. The broad any-pre-onset definition is useful for assessing patient-level enrichment, but it can include alerts much earlier than the nominal 6-hour target. Under the stricter horizon-concordant 0–6 hour definition, the top-2% policy detected 17.8% of sepsis stays with a median lead time of 2.34 hours. Under a broader but clinically plausible 0–24 hour early-warning definition, detection was 29.3% with a median lead time of 4.87 hours. Thus, the alarm analysis should be read as a family of operating definitions rather than a single detection number. The any-pre-onset result is secondary and should not be described as pure 6-hour horizon performance.

Clinical meaning of the leakage-ablation result

The CLEAN versus FULL comparison is one of the central findings of the study. In many EHR models, missingness and measurement recency are not merely technical nuisances; they can encode clinical suspicion. A newly obtained lactate, coagulation test, culture, or arterial blood gas often means that a clinician is already worried. If such process information is used without qualification, the model may appear predictive while functioning as a detector of clinician concern. The present CLEAN model intentionally removes measurement-age and observation-mask features from the primary model. The minimal performance loss suggests that the remaining physiologic signals carry genuine predictive information.

This finding has two implications. First, it strengthens the scientific validity of the reported SHAP explanations, because the top variables are actual physiology and temporal trends rather than test-ordering artifacts. Second, it makes the feature-set rationale more transparent. The study can report that a more permissive feature set was tested, that suspected leakage channels were removed, and that the core conclusions did not depend on those channels. This directly addresses concerns raised in systematic reviews about heterogeneity, incomplete reporting, and limited reproducibility in sepsis prediction studies (Islam *et al.* 2023; Zubair *et al.* 2025).

Residual missingness leakage cannot be fully excluded, because native LightGBM split directions can still learn whether a nu-

meric value is missing even when explicit observation masks and measurement-age variables are removed. The median-imputation sensitivity analysis addresses this concern directly: forcing numeric missing values to training-set medians reduced 6-hour AUROC to 0.766. This modest drop relative to the no-year primary model (0.772) suggests that implicit missingness contributes some signal, but the model's discrimination is not predominantly a missingness artifact. Similarly, removing lactate-derived features left 6-hour AUROC at 0.776, reducing concern that performance is driven by lactate ordering or lactate-specific suspicion pathways.

Discussion of global SHAP findings

The global SHAP profile is consistent with prior interpretable sepsis studies. Glasgow Coma Scale total score was the strongest global feature in the 6-hour model. This aligns with Hu *et al.* (2022), who also identified GCS as the most important feature in an interpretable MIMIC-IV sepsis mortality model. Although our endpoint is future Sepsis-3 onset rather than mortality, severe neurologic depression is a marker of systemic illness, hypoperfusion, encephalopathy, sedation burden, or respiratory failure. A low GCS may therefore signal either organ dysfunction already emerging or vulnerability to rapid deterioration.

Blood urea nitrogen was another major contributor, and its dependence plot showed increasing SHAP contribution at higher values. This agrees with Hu *et al.* (2022), who identified blood urea nitrogen among the top predictors of sepsis prognosis. BUN is not a pure sepsis biomarker; it reflects renal perfusion, catabolic stress, dehydration, renal dysfunction, and critical illness severity. In our model, BUN likely contributes because renal dysfunction is both a component of organ failure and a marker of systemic deterioration. The fact that serum creatinine change and urine-output dynamics also appeared among important predictors strengthens the interpretation that renal physiology is a key pre-onset signal.

Urine-output dynamics were especially informative. The range of 6-hour urine output over the previous 12 hours ranked third globally and was one of the top numeric dependence features. Low recent urine-output range increased predicted risk, while high variability had lower SHAP contribution. This is consistent with the role of urine output in SOFA renal scoring and with prior studies identifying urine output as important in sepsis prognosis (Hu *et al.* 2022; Zhang *et al.* 2025). Clinically, persistently low or nonrecovering urine output can reflect renal hypoperfusion, shock physiology, fluid balance abnormalities, or early acute kidney injury. A dynamic urine feature is therefore more informative than a single static urine value because it captures trajectory.

The urine-output features also required explicit outlier scrutiny. The raw landmark table contained implausibly high urine-derived values, with 6-hour urine output maxima above physiologic ranges and 0.7% of 12-hour urine-range landmarks exceeding 5,000 mL. Such values may reflect charting artifacts, duplicate aggregation, GU irrigation documentation, or non-urine output events mapped into structured output streams. For this reason, dependence plots were clipped to the 1st–99th percentile and a winsorized sensitivity analysis was added. Clipping urine-related features at training-set 1st–99th percentile limits produced h=6 AUROC 0.774 versus 0.772 for the no-year primary model, indicating that the main conclusion is not driven by extreme urine-output values. Future external implementations should apply site-specific urine-output plausibility filters before deployment.

The overlap between predictors and the Sepsis-3 label requires careful interpretation. Several important variables, including GCS, platelet count, bilirubin, creatinine, urine output, oxygenation-related proxies, and blood pressure, are direct SOFA components or close physiologic relatives of SOFA components. The model is therefore best described as forecasting future EHR-defined Sepsis-3 onset, partly by detecting early movement toward an organ-dysfunction threshold. This is clinically useful for early warning, but it is not the same as identifying a latent biological sepsis state independent of the label construction. The no-direct-SOFA-component sensitivity model reduced 6-hour AUROC to 0.758, indicating that SOFA-related physiology contributes meaningfully but does not solely explain the predictive signal.

Body temperature was the second most important global feature and had a nonlinear dependence pattern. Normothermic values around 36.5–37.0 °C were generally protective, while higher values increased risk. This is consistent with infection physiology and with Zhang *et al.* (2025), who identified temperature among the top features for multiple sepsis prognosis prediction in MIMIC-IV. The temperature feature was also technically strengthened by harmonizing Fahrenheit and Celsius charting; without that correction, temperature observation would have been artificially sparse and less reliable. The dependence curve should still be interpreted cautiously, because temperature can be affected by antipyretics, external warming/cooling, perioperative state, and measurement method.

Inflammatory and perfusion-related variables also behaved as expected. White blood cell count was ranked seventh globally and was a major positive contributor in the local SHAP example. Lactate itself was not in the global top 10, but lactate range over 12 hours and local lactate value contributed meaningfully. This is a useful distinction: lactate is often measured when clinicians are already worried, which is why measurement-process variables around lactate were excluded from CLEAN. The retained lactate value and lactate dynamics are physiologic variables; when present, they still provide useful signal. In the local case, lactate 3.5 mmol/L, WBC 20.2 K/uL, systolic blood pressure 88 mmHg, and GCS 3 formed a high-risk physiologic pattern.

Admission type remained influential in the primary no-year SHAP analysis, while deidentified admission-year group appeared only in the year-including internal benchmark. These variables should not be interpreted as direct causal mechanisms. Hospital admission type likely captures case-mix, route of presentation, acuity, and workflow differences between emergency, urgent, elective, observation, and surgical admissions. Anchor year group may capture secular changes in charting, coding, ICU practice, patient mix, or COVID-era effects; for that reason, it was removed from the primary deployment-oriented model and retained only as an internal benchmark. Prior MIMIC-IV fairness work has shown that demographic and administrative variables can reveal performance and fairness concerns in clinical models (Meng *et al.* 2022). For that reason, admission route and other administrative variables should be examined in subgroup calibration and fairness analyses before any clinical use.

The primary no-year model and temporal validation analyses address this concern directly. The year-including internal benchmark reached 6-hour AUROC 0.779, while removing deidentified admission-year group produced AUROC 0.772. Temporal validation, which trained on earlier deidentified year groups and tested on 2020–2022 without year group as a feature, produced 6-hour AUROC 0.762 and AUPRC 0.047. This temporal drop is expected under calendar-period shift and reinforces that the present model should be presented as a single-database benchmark requiring external validation, not as a transportable deployment model.

Local explanation and patient-level plausibility

The selected local SHAP case illustrates why patient-level explanations are useful. At ICU hour 8, the no-year model estimated a 45.2% probability of Sepsis-3 onset within the next 6 hours, and onset occurred 4.35 hours later. The largest positive driver was GCS 3, followed by hemoglobin, fever-range temperature, lactate, lactate range, emergency admission type, leukocytosis, low urine-output dynamics, hyperglycemia, bilirubin, tachycardia, and respiratory-rate features. This explanation does not merely list features; it describes a recognizable clinical syndrome of neurologic depression, inflammatory response, hypoperfusion, and early organ dysfunction.

The local explanation also reveals the difference between global and individual importance. Some variables that were not dominant globally, such as hemoglobin, glucose, bilirubin, or systolic blood pressure, contributed materially in this patient. These local attributions are context-dependent and can reflect interactions or correlated severity markers rather than isolated causal effects. This supports the use of both global and local SHAP in the manuscript. Global SHAP answers which variables matter on average; local SHAP answers why a specific patient was flagged at a specific time. In high-stakes early-warning settings, both views are necessary because clinicians need population-level trust and case-level interpretability.

Implications for clinical translation

The alarm-policy results should be interpreted as operating-point analysis, not as proof of clinical utility. The top-2% h=6 no-year policy gives a transparent trade-off: 17.8% horizon-concordant detection, 29.3% 0–24 hour detection, and 40.3% any-pre-onset enrichment at 15.5 alerts per 100 patient-days. It also requires about 10.5 emitted alerts per horizon-concordant detected case, which is a substantial operational cost. Prospective studies are needed to determine whether such alerts change clinician behavior, shorten time to antibiotics or cultures, reduce organ dysfunction, or improve mortality. The COMPOSER deployment study is an important reference point because it evaluated process and outcome changes after implementation rather than relying only on retrospective discrimination (Boussina *et al.* 2024).

The current model is intentionally simpler than recent multimodal or language-model systems (Shashikumar *et al.* 2025; Yang *et al.* 2026). A tabular LightGBM model can be trained reproducibly, explained with Tree SHAP, audited for leakage, and integrated into a landmark-alarm analysis with relatively low computational overhead. More complex models may improve discrimination by using notes, semantic abstraction, waveforms, or multimodal context, but they also introduce additional challenges in interpretability, maintenance, data availability, and external validation. The present work can therefore serve as a transparent baseline before moving to richer models.

Limitations

Several limitations should be acknowledged. First, this is a single-database retrospective study using MIMIC-IV v3.1. Although MIMIC-IV is widely used and carefully deidentified, it reflects one health system and its documentation practices. The added temporal validation reduces, but does not eliminate, concerns about transportability. External validation on other ICU databases is required before generalizing the operating thresholds. Second, Sepsis-3 onset labels are EHR-derived and depend on antibiotic, culture, and SOFA reconstruction rules. Misclassification is possible when infection suspicion is not captured by available medication or microbiology records, when cultures are delayed, or when organ dysfunction is incompletely observed.

Third, the forecastable cohort excludes Sepsis-3 onset within the first 4 ICU hours; therefore, the model is not intended to detect sepsis already present at ICU arrival. Fourth, death or discharge before a horizon is treated as censoring for the primary binary sepsis-onset task. This avoids labeling unobserved futures as non-sepsis, but death without observed sepsis can also be viewed as a competing event. A 6-hour composite sepsis-or-death sensitivity model achieved AUROC 0.808 and AUPRC 0.140, suggesting that broader deterioration is more discriminable than sepsis onset alone, but it is a different clinical endpoint. Fifth, the model uses structured data only. Clinical notes, imaging reports, waveform data, bedside nursing notes, and microbiology text could contain additional early information. Recent LLM-based and multimodal studies suggest that unstructured context can improve sepsis prediction in uncertain cases (Shashikumar *et al.* 2025; Yang *et al.* 2026).

Sixth, the CLEAN feature set deliberately removes measurement-process channels and vasopressor exposure to reduce leakage, but this conservative choice may also remove clinically useful real-time information. In deployment, one could evaluate multiple policies: a physiology-only warning model for early anticipation and a workflow-aware model for situational awareness. Seventh, SHAP explanations describe model behavior rather than causal effects. A high SHAP value for BUN, temperature, admission type, or urine-output range does not prove that changing that variable will change the patient's risk. The explanations should be interpreted as attribution within the fitted model. Eighth, alarm-policy performance was evaluated retrospectively. The false-alarm rate, clinician response, and patient benefit may change when alerts are displayed in real time. A prospective silent trial followed by pragmatic implementation would be required before clinical claims.

Future works

Future work should proceed in four directions. First, external validation should test the CLEAN model and alarm thresholds on independent ICU cohorts, ideally including eICU, AmsterdamUMCdb, or other contemporary ICU datasets. Second, subgroup performance should be evaluated by sex, age group, admission type, anchor year group, ICU type, and baseline severity to assess calibration and fairness. Third, trajectory-aware models such as recurrent neural networks, temporal convolutional networks, transformers, or semantic-abstraction LLM pipelines should be compared against the CLEAN LightGBM baseline under the same leakage-audited landmark labels. Fourth, a silent prospective evaluation should measure alert frequency, lead time, clinician agreement, and whether alerts would trigger clinically appropriate review before any interventional deployment.

Overall, the study supports a pragmatic conclusion: a carefully engineered, leakage-audited LightGBM model can provide interpretable Sepsis-3 onset forecasts from routine ICU physiology. Its discrimination is not as high as mortality prediction models, but that is expected because the task is earlier, rarer, and more clinically prospective. The strongest scientific contribution is the combination of dynamic multi-horizon labels, CLEAN feature auditing, alarm-policy evaluation, and SHAP-based physiologic interpretation. This combination makes the model a transparent benchmark and a foundation for future externally validated early-warning research.

CONCLUSION

This study developed and evaluated a leakage-audited, interpretable machine-learning pipeline for forecasting future Sepsis-3 onset in adult ICU patients using MIMIC-IV v3.1. Rather than treating sepsis prediction as a single static classification task, the analysis formulated prediction as a repeated landmark problem: at each ICU hour, the model estimated whether Sepsis-3 onset would occur within clinically relevant future windows of 4, 6, 12, and 24 hours. This design allowed the model to represent the dynamic nature of critical illness while preserving patient-level separation between training, validation, and test cohorts.

The main empirical finding is that routinely collected ICU physiology contains signal for future EHR-defined Sepsis-3 onset. The no-year CLEAN LightGBM model achieved test AUROCs of 0.782, 0.772, 0.750, and 0.720 across the 4-, 6-, 12-, and 24-hour horizons, respectively. The 6-hour horizon provided the most clinically balanced operating point, but alarm interpretation depends on the detection window. Under a top-2% alarm policy with a 6-hour cooldown, the model detected 17.8% of sepsis cases within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition, with an alarm burden of 15.5 alerts per 100 patient-days.

A central contribution of the study is the explicit separation of physiologic prediction from workflow-driven leakage. Removing vasopressor exposure, measurement-recency variables, and observation-mask channels caused only minimal performance degradation, with AUROC decreases below 0.01 across all horizons. This result suggests that model performance was not primarily driven by clinician test-ordering behavior or treatment proxies. For academic and clinical interpretation, this is important because it strengthens the claim that the model learned from patient state and physiologic trajectories rather than from retrospective artifacts of care delivery.

The explainability analyses further support physiologic face validity. Global SHAP analysis identified neurologic status, body temperature, urine-output dynamics, blood urea nitrogen, inflammatory markers, vital signs, and acid-base or lactate-related trends as major contributors. These variables align with known domains of Sepsis-3 organ dysfunction and with prior interpretable sepsis studies. Local SHAP analysis in an example patient showed that the model's high-risk prediction was driven by depressed consciousness, fever-range temperature, elevated lactate, leukocytosis, borderline hypotension, abnormal renal-output dynamics, and emergency admission context. Dependence plots showed nonlinear effects for the most influential numeric predictors.

Taken together, these findings support the proposed pipeline as an internally and temporally validated benchmark for Sepsis-3 onset forecasting. The model is not presented as a clinical decision-support system; it remains a retrospective, single-database study requiring external validation, prospective silent evaluation, subgroup calibration monitoring, and workflow testing. Nevertheless, the combination of dynamic horizon-specific labels, conservative feature auditing, interpretable LightGBM modeling, SHAP-based explanation, and alarm-policy evaluation provides a reproducible framework for future early-warning research. The results indicate that a physiology-based model can identify a subset of ICU patients at increased risk of future EHR-defined Sepsis-3 onset while producing explanations that can be inspected at both cohort and patient levels.

Supplementary Material

The submission package includes supplementary CSV and Markdown files containing baseline characteristics, patient-level split counts, label observability by split and horizon, the patient-level bootstrap table, decile calibration, temporal-validation results with bootstrap intervals, temporal alarm-policy results, clinical-score and calibrated logistic-regression baselines, sensitivity analyses, revised alarm-window metrics with bootstrap intervals, subgroup performance and subgroup alarm summaries, sepsis-onset timing detection, no-year SHAP summaries, categorical SHAP summaries, urine-output plausibility checks, competing-risk sensitivity results, a related-work comparison table, a CLEAN feature dictionary with source tables and MIMIC item identifiers, package versions, a TRIPOD-AI/CLAIM-oriented reporting checklist, a synthetic landmark-labeling smoke-test example, a reference-metadata spot-check note, and a reproducibility note documenting landmark generation, labeling, model settings, class-imbalance handling, and censoring rules.

Availability of data and material

This study uses MIMIC-IV v3.1, which is available through PhysioNet to credentialed users who complete the required data-use training and data-use agreement. Code for cohort construction, feature generation, landmark labeling, model training, evaluation, and figure generation will be released in a public repository upon acceptance. An anonymized code archive will be made available to reviewers as confidential supplementary material where permitted by journal policy. The repository will not contain MIMIC-IV data and must be run by credentialed users with local MIMIC-IV access. The supplementary material lists the derived feature dictionary, item identifiers, package versions, random seed, model hyperparameters, labeling rules, and a synthetic smoke-test example needed to verify the labeling logic without access to protected clinical data.

Acknowledgements

The authors would like to thank researcher who publicly share their dataset in kaggle. The authors acknowledge Sakarya University of Applied Sciences (https://subu.edu.tr/) for the technical support provided to publish the present manuscript.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Boussina, A., S. P. Shashikumar, A. Malhotra, R. L. Owens, R. El-Kareh, *et al.*, 2024 Impact of a deep learning sepsis prediction model on quality of care and survival. npj Digital Medicine **7**: 14.

Duval, L., A. Villié, F. Zheng, G. Terraz, S. Blein, *et al.*, 2025 Early prediction of vasopressor initiation in ICU sepsis patients using an interpretable EHR-based ML model. BMC Medical Informatics and Decision Making **25**: 442.

Gao, J., Y. Lu, N. Ashrafi, I. Domingo, K. Alaei, *et al.*, 2024a Prediction of sepsis mortality in ICU patients using machine learning methods. BMC Medical Informatics and Decision Making **24**: 228.

Gao, Y., C. Wang, J. Shen, Z. Wang, Y. Liu, *et al.*, 2024b Systematic review and network meta-analysis of machine learning algorithms in sepsis prediction. Expert Systems with Applications **245**: 122982.

Hu, C., L. Li, W. Huang, T. Wu, Q. Xu, *et al.*, 2022 Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infectious Diseases and Therapy **11**: 1117–1132.

Islam, K. R., J. Prithula, J. Kumar, T. L. Tan, M. B. I. Reaz, *et al.*, 2023 Machine learning-based early prediction of sepsis using electronic health records: a systematic review. Journal of Clinical Medicine **12**: 5658.

Johnson, A., L. Bulgarelli, T. Pollard, B. Gow, B. Moody, *et al.*, 2024 MIMIC-IV. Version 3.1.

Johnson, A. E. W., L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, *et al.*, 2023 MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**: 1.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, *et al.*, 2017 LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154.

Liu, Z., W. Shu, T. Li, X. Zhang, and W. Chong, 2025 Interpretable machine learning for predicting sepsis risk in emergency triage patients. Scientific Reports **15**: 887.

Lu, X., H. Kang, D. Zhou, and Q. Li, 2022 Prediction and risk assessment of sepsis-associated encephalopathy in ICU based on interpretable machine learning. Scientific Reports **12**: 22621.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, *et al.*, 2020 From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence **2**: 56–67.

Meng, C., L. Trinh, N. Xu, J. Enouen, and Y. Liu, 2022 Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Scientific Reports **12**: 7166.

Seymour, C. W., V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, *et al.*, 2016 Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA **315**: 762–774.

Shan, W., D. Sun, and Z.-P. Liu, 2024 Predicting sepsis onset in ICU patients using machine learning and feature section: a case study of MIMIC-IV data. In *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pp. 546–551.

Shashikumar, S. P., S. Mohammadi, R. Krishnamoorthy, A. Patel, G. Wardi, *et al.*, 2025 Development and prospective implementation of a large language model based system for early sepsis prediction. npj Digital Medicine **8**: 290.

Singer, M., C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, *et al.*, 2016 The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA **315**: 801–810.

Yang, Y., B. Yi, and Y. Chen, 2026 Lightweight large language models for early sepsis prediction via a semantic abstraction rule engine. Scientific Reports .

Zhang, S.-Z., H.-Y. Ding, Y.-M. Shen, B. Shao, Y.-Y. Gu, *et al.*, 2025 Harness machine learning for multiple prognoses prediction in sepsis patients: evidence from the MIMIC-IV database. BMC Medical Informatics and Decision Making **25**: 152.

Zhang, Y., T. Kirchler, and A. P. Wang, 2026 Evaluating artificial intelligence for sepsis prediction in emergency departments: a systematic review and meta analysis. Journal of Medical Systems **50**: 49.

Zubair, M., I. Din, N. Sarwar, B. Elov, and Z. Trabelsi, 2025 Revolutionizing sepsis diagnosis using machine learning and deep learning models: a systematic literature review. BMC Infectious Diseases **25**: 1396.

How to cite this article: Mansour, M., and Donmez T. B. An Explainable Leakage-Audited Framework for Multi-Horizon Sepsis-3 Onset Forecasting from Complex ICU Dynamics. *Chaos and Fractals*, 3(2), 76-91, 2026.

ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.92-107
https://doi.org/10.69882/adba.chf.2026073

Integrating Fuzzy Logic and Fractional-Order Operators in Convolutional Neural Networks for Handwritten Digits and Characters Recognition

Akshara Sreenivasan [*,1], **Vinodkumar Arumugam** [*,2], **Sriramakrishnan Pathmanaban** [*,3], **Saravanan Arumugam** [β,4] and **Jehad Alzabut** [α,γ,5]

*Department of Mathematics, Amrita School of Physical Sciences, Coimbatore, Amrita Vishwa Vidyapeetham, India , βDepartment of Mathematics, Sona college of Technology, Salem, 636 005, Tamilnadu, India , αDepartment of Mathematics and Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia, γDepartment of Industrial Engineering, OSTIM Technical University, Ankara 06374, Türkiye.

ABSTRACT Handwritten character recognition is vital for document digitization and autonomous reading systems. This study focuses on enhancing handwritten Devanagari character recognition using deep learning models, specifically fine-tuned Convolutional Neural Networks (CNNs), combined with hybrid mathematical methods for image enhancement. Devanagari script, used for several South Asian languages, presents challenges due to its complexity and structural variations. To improve recognition accuracy, we propose a fuzzy-enabled Power-Law transformation for image enhancement, along with other techniques like Grunwald-Letnikov Fractional Differentiation (GLFD) and Atangana-Baleanu-Riemann (ABR). Experimental results show that the CNN model with Power-Law+ Fuzzy enhancement achieves the highest accuracy (98.02%), surpassing even MobileNetV2 (92.86%). The same method also yields impressive performance on the MNIST dataset (99.15% accuracy), demonstrating its effectiveness across different scripts. These findings highlight the benefits of integrating advanced preprocessing with deep learning for improved handwriting recognition, offering practical applications in multilingual document processing and OCR-based automation.

KEYWORDS
Handwritten character recognition
Devanagari script
Image enhancement techniques
Convolutional neural networks (CNNs)
Fractional-order operators

INTRODUCTION

In numerous practical applications, recognizing handwritten scripts is a key task, particularly for scripts that feature intricate characters like Devanagari. A variety of prominent Indian languages, such as Hindi, Sanskrit, Marathi, Konkani, and Nepali, utilize the Devanagari script. One of its 48 fundamental characters, the shirorekha, is a defining horizontal line that runs along the top of complete letters, and it consists of 14 vowels and 34 consonants. The writing direction is from left to right (Jayadevan *et al.* 2011). Moreover, the addition of diacritical signs (matras), which represent vowel sounds, might expand the set of acceptable character forms. Consonants are used to create compound characters, or ligatures, which add to the script's complexity and result in an extensive character set and tremendous structural variety (Pal and Chaudhuri 2004).

Reliable handwritten Devanagari writing recognition can improve the handling of these scripts by natural language processing systems, automate form reading, and digitize archives. This is quite difficult since the script has many complex components, including combined letters (conjuncts), modifiers, many strokes, and peculiar lines (shirorekha) (Arora *et al.* 2024). In addition, a character may be written by the same person at different times in different sizes, shapes, or orientations because everyone writes differently. Other issues, such as differences in print thickness, noise, or scanning distortions, make recognition even more difficult. For optical character recognition (OCR) to recognize Devanagari text, the image must first be cleaned up (preprocessing), then it must be segmented into words, lines, and characters (segmentation), identify the salient features of each character, classify them (often using

Manuscript received: 28 March 2026,
Revised: 29 June 2026,
Accepted: 4 July 2026.

[1]cb.sc.i5das20139@cb.students.amrita.edu,
[2]a_vinodkumar@cb.amrita.edu.(**Corresponding author**)
[3] p_sriramakrishnan@cb.amrita.edu
[4]mathsaran@gmail.com
[5]jalzabut@psu.edu.sa

machine learning), and fix any errors (posting). Character separation and feature extraction are very difficult tasks in Devanagari OCR because of the complexity of the script.

To improve generalization, besides the Devanagari script, we also applied our methods to the Modified National Institute of Standards and Technology (MNIST) dataset, which is a well-known benchmark for handwritten digit recognition. The MNIST dataset comprises 70,000 grayscale images of handwritten digits (0–9), with each image normalized to 28×28 pixels. First published by LeCun et al., this dataset has since become a de facto standard testbed for evaluating machine learning algorithms in image classification (LeCun *et al.* 1998). Although MNIST digits are more structurally primitive than Devanagari letters, they too pose considerable difficulty because of variability in writing style, stroke width, slant angles, and placement. While these variations are, in fact, simpler than those found in Devanagari, we believe that the standardization known to exist within MNIST makes it an excellent benchmark to test the overall effectiveness of our techniques designed to improve performance on diverse writing systems. Testing our techniques on the intricately detailed Devanagari script, along with the more thoroughly researched MNIST dataset, allows us to test the resilience and flexibility of our method across various recognition problems.

Modern OCR systems prefer to use convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformer-based models that learn end-to-end from raw image inputs without using handcrafted features (Khaparde and Mankar 2018). Integration with natural language models also improves accuracy by placing context on ambiguous character predictions. Image enhancement plays a notable role in deep learning-based OCR classification and contributes significantly to recognition accuracy. Normal preprocessing steps, such as binarization, noise reduction, deskewing, and contrast correction, help in restraining gradient changes before interfacing with the neural network so that the models can focus on salient features instead of irrelevant image quality discrepancies. High-stabilized feature extraction, reduced fraction of needed training time, and improved generalization in documents of diverse types are observed due to enhanced images. Image enhancement is particularly critical for reliably recognizing poor-quality historical documents, worn receipts, or low-resolution scans.

Prominent other modern OCR systems which implement CNNs are Tesseract v4, which switched from LSTM-based neural networks to Tesseract's LSTM-based neural networks, Google Cloud Vision OCR, and Microsoft Computer Vision API. Overall, these algorithms use CNNs for recognition, as well as attend to character shape patterns that can be exploited as features. This problem is well suited to the operations of CNNs, which learn hierarchical levels of abstraction using convolutional layer frameworks, simple shallow edges in the early layers and progressively intricate character motifs in deeper layers.

Research Gaps and Challenges

Due to the variability and noise in handwritten samples, the recognition accuracy of traditional handwritten character recognition systems for the Devanagari script is low. Blurry scans of documents can capture images of characters that are faint, poorly penned, and disjointed. Exceedingly large datasets containing numerous authors and differing conditions add complexity to the task, and conventional methods of pattern recognition are incapable of dealing with such diversity. Rugged and outline-based approaches, which rely on identified features, as well as template-based methods, due to their lack of flexibility and adaptability, do not stand a chance in more complex situations. The primary obstacle is precisely distinguishing Devanagari characters amidst differing writing styles, stroke thicknesses, angles, and obstructions. Without additional image polishing, machine learning approaches risk classifying unconventional handwriting or low contrast characters as noise. Thus, advanced preprocessing techniques coupled with modern deep learning frameworks stand to improve image quality and recognition performance.

Moreover, most existing methods for script recognition do not effectively recognize handwritten numerals. This work uses deep learning architectures along with image enhancement techniques to automate the detection of variability and noise in digits and the Devanagari characters.

Research Highlights and Contributions

To address the research gap discussed earlier, developing a robust handwritten character recognizer for the Devanagari script would significantly advance information retrieval and cultural heritage preservation by converting analog text into editable, searchable digital formats. This has direct implications for libraries, government digitization initiatives, educational content processing, and accessibility software.

In this work, we propose using a power-law transformation, which employs an exponential function to modify pixel intensities (Gonzalez and Woods 2002). This approach improves the average brightness and contrast of an image by redistributing intensity values. A gamma value of less than one makes dark regions lighter, which can enhance elusive handwriting or thin strokes. On the other hand, gamma values greater than one will reduce exposure to bright regions, thus making some subtle details more pronounced. Images with low contrast and poor lighting are best served with this technique, as through a delicate balance of the transformation parameters, faint strokes in handwritten characters that are weak or faded can be rendered visible and decipherable. With the process completed, the images undergo classification and comparison using two deep learning models: a custom CNN and MobileNetV2. To ensure model stability, dataset ablation techniques were applied using the MNIST dataset.

This study has the following key contributions:

- We design a fuzzy-enhanced Power-Law transformation specifically aimed at improving images for Devanagari character recognition.
- We study the impact of other fuzzy image enhancement techniques with Grunwald-Letnikov Fractional Differentiation (GLFD) and Atangana-Baleanu-Riemann (ABR), on the recognition accuracy of handwritten Devanagari characters.
- With a custom CNN and MobileNetV2 architecture, we apply the proposed image enhancement techniques and analyze the response of the models to enhanced inputs.
- We show the application of preprocessing techniques in handwriting recognition systems and how they can be used to improve classification performance.
- In testing the generalization capabilities of our models, we perform additional benchmark testing using the MNIST dataset.

The rest of the manuscript is organized in this manner: Section 2 examines the current literature on image enhancement methods, cutting-edge deep learning models, and recent progress in handwritten character recognition. Section 3 outlines the datasets and assessment criteria utilized in our experiments, featuring the Devanagari dataset and the MNIST dataset. Section 4 outlines our

approach, including preprocessing procedures, the suggested image enhancement methods, and the deep learning models utilized. Section 5 displays the experimental findings and evaluation for both the Devanagari and MNIST datasets, emphasizing the influence of preprocessing on recognition precision. Section 6 examines the constraints of our method and highlights possible avenues for future investigation. Section 7 concludes the paper by recapping our results and exploring their significance for systems that recognize handwritten characters.

RELATED WORK

This section reviews existing image enhancement techniques, discusses available deep learning models, and summarizes recent advancements in handwritten character recognition.

Convention Image Enhancement Techniques

Over the past few decades, image enhancement has played a crucial role in classification and recognition tasks, particularly with the advent of numerous deep learning models (Krizhevsky *et al.* 2017). These techniques are applied in image processing to improve image quality and emphasize significant features, thereby making images more suitable for analysis or recognition by deep learning algorithms. In the context of handwritten character recognition, image enhancement can be especially vital, as even minor improvements in clarity and sharpness can significantly enhance a model's ability to differentiate between similar characters (Garg and Singh 2012; Sharma and Kumar 2021). Here, we summarize two enhancement methods that are relevant to this work:

GLFD is an extension of traditional differentiation concepts in mathematics, utilizing non-integer orders of differentiation. This method acts as an advanced edge detection filter in imaging (Podlubny 1999). While classical first-order derivatives enhance edges by detecting sharp intensity changes, fractional-order derivatives (with orders between 0 and 1) provide greater flexibility, allowing for more effective enhancement of fine details. The Grunwald–Letnikov method calculates a fractional difference for each pixel, effectively highlighting medium and high-frequency components (such as edges and textures) without eliminating low-frequency (smooth) components. In the case of character images, GLFD has been used to accentuate edges and highlight subtle stroke textures that might otherwise go unnoticed with conventional filters. This enhancement can make delicate curvatures and line tips of handwritten characters more pronounced, potentially enabling recognition models to distinguish characters with near-similar appearances more easily (Das and Gupta 2013).

The ABR technique is designed to make low-contrast photos clearer. It's especially good at highlighting differences in both small areas and the overall image (Atangana and Baleanu 2016). Its strength comes from combining fractal geometry and fractional calculus, which helps it capture fine details without losing important parts of the image, like edges and textures. The ABR method makes sure that subtle changes in the image are emphasized while keeping key structural elements intact by using these mathematical ideas. Both methods help improve image quality, especially when recognizing handwritten characters. ABR is better at preserving texture and increasing contrast, while GLFD focuses on sharpening edges and small details to bring out subtle features more clearly. Together, they create a strong approach for preparing images, making them better suited for deep learning-based recognition systems.

Deep Learning for Character Recognition: Deep learning has revolutionized character recognition by directly extracting important features from raw image data, eliminating the need for time-consuming manual feature engineering. Two well-known types of deep neural networks that are particularly useful in this context are Convolutional Neural Networks (CNNs) and MobileNetV2.

CNNs (Convolutional Neural Networks) have become the preferred choice for image recognition tasks, including Optical Character Recognition (OCR). These networks use multiple layers of convolution filters to process images and detect features like edges, curves, or more complex shapes as they move through deeper layers. Each convolution layer uses a set of learned kernels that activate when specific patterns in the image are detected (LeCun *et al.* 1998). Following certain convolution layers, pooling layers reduce the spatial resolution of feature maps while retaining the most critical information. By stacking convolution and pooling layers, CNNs build a hierarchical representation of features starting with simple edges in early layers and progressing to complex character shapes in later layers. At the output, fully connected (dense) layers utilize these extracted features to classify the image into one of the predefined character classes. CNNs are especially well-suited for character recognition because their learned features are translation-invariant, allowing them to handle minor variations and distortions in handwriting effectively. Over the past decade, CNN-based methods have consistently outperformed traditional approaches for handwritten character recognition across various scripts.

MobileNetV2 is a specialized CNN architecture designed for mobile and embedded systems (Sandler *et al.* 2018). It is a lightweight network that achieves high performance with significantly fewer parameters and computational operations compared to larger models. The key innovation in MobileNetV2 is the use of depth-wise separable convolutions, which break down a standard convolution into two steps: first, a depth-wise convolution processes each input channel independently, and second, a point-wise (1×1) convolution combines the outputs across channels. MobileNetV2 reduces computation while still learning important features. It uses an inverted residual design with linear bottlenecks: each block expands the channels, applies depth-wise convolution, and then compresses them again. Linear bottlenecks preserve information, and shortcuts between blocks improve gradient flow and reuse features. After training on large datasets like ImageNet, it can be fine-tuned for character recognition, offering high accuracy with low computational needs and making it ideal for handwriting recognition on mobile or low-power devices (Howard *et al.* 2017).

Works with Handwritten Devanagari Characters Recognition

Several strategies for detecting handwritten Devanagari characters have been developed over the years. Early techniques relied on typical machine learning classifiers with manually produced features. Researchers used approaches such as SVMs and k-NN to extract features like zoning characteristics or histograms of directed gradients before training classifiers to recognize characters. While moderately effective, these approaches suffered from handwriting variations and noise, particularly as datasets rose in size and diversity. As a result, they frequently failed to scale effectively with rising data complexity (Sethi and Singh 2020).

Deep learning approaches have gained popularity in recent years due to their ability to learn discriminative features directly from raw data without the need for manual feature engineering. The performance of large-scale CNN-based models has outperformed traditional methods when applied to Devanagari character

datasets. These models work very well because human writing naturally varies in handwriting styles and stroke shapes. For example, a study on deep CNNs for large-scale Devanagari character recognition showed that by internally capturing minute aspects of the script, deeper networks could maintain good performance even when the dataset size expanded considerably. Researchers have investigated mixing deep learning with picture preprocessing methods to further increase recognition accuracy (Agrawal and Nair 2019). One such approach is to use fractional differentiation methods, such as the Generalized Local Fractional Derivative (GLFD), before CNN classification. By sharpening the input images with GLFD, studies were able to improve the accuracy of feature extraction and recognition. By highlighting subtle differences in stroke patterns, the sharper images helped CNNs understand character shapes, particularly when it comes to recognizing characters with similar outlines. Another innovative preprocessing method is the application of ABR (Atangana-Baleanu-Reimann) fractal-fractional differentiation techniques. These methods use fuzzy principles to improve local contrast and reduce noise, resulting in crisper handwritten Devanagari characters. Training CNNs on these updated images boosted the models' ability to distinguish different features. Combining deep learning and excellent preprocessing has considerably improved the accuracy of handwritten Devanagari OCR systems, outperforming any single method.

Traditional machine learning classifiers with manually constructed features were employed in early research on Devanagari character detection. However, accuracy and scalability have significantly increased due to recent developments in deep learning, particularly CNNs. A potent method for improving handwritten Devanagari character recognition is the combination of deep learning models with sophisticated preprocessing techniques.

Works with MNIST Digits Recognition

The industry standard for assessing machine learning techniques has long been the MNIST dataset. The accuracy rates of the first methods, which used handcrafted features and conventional classifiers like SVMs and k-NN, were 95–98% (Liu *et al.* 2003). Model robustness and efficiency can still be tested with MNIST, even if it is regarded as a solved problem. The goal of recent research is to maintain good performance while minimizing model size and computing demands, particularly for devices with limited resources (Sandler *et al.* 2018; Howard *et al.* 2017).

In contrast to scripts like Devanagari, image enhancement techniques for MNIST have received relatively less attention because modern CNNs already perform exceptionally well on this benchmark dataset. However, experiments conducted under degraded conditions, such as noise, blur, or low contrast, have shown that effective preprocessing can significantly boost performance. Through the comparison of various improvement techniques on MNIST and other datasets, such as Devanagari, researchers may learn more about how the advantages of preprocessing vary depending on the dataset's characteristics and character complexity. This comparative study clarifies how preprocessing improves model performance at different data quality and script complexity levels.

MATERIALS AND METRICS

This section outlines the dataset sources and performance metrics employed to evaluate the models for handwritten character identification.

Materials

We use a sizable publicly accessible dataset of handwritten Devanagari characters for this study. Over 92,000 grayscale pictures of isolated character glyphs make up this collection, which represents 36 different classes of the Devanagari alphabet (one class per unique letter). The UCI Machine Learning Repository makes the dataset available (Acharya and Gyawali 2015). The collection's images all have a single handwritten character on a simple backdrop. The samples, which cover a broad spectrum of handwriting styles, were gathered from several writers and sources. Variations include variations in slant, where some authors lean to the right or left when writing, and variations in stroke thickness, with some characters being created with thinner pens and others are using thicker markers. Additionally, the images contain imperfections such as faint smudges of ink, background speckles, or uneven scanning lighting, which contribute to their realism.

The Modified National Institute of Standards and Technology (MNIST) dataset serves as a standard yardstick for concept categorization and machine intelligence tasks. It consists of 70,000 in manuscript number concepts (0–9), each accompanying a resolution of 28×28 pixels (LeCun *et al.* 2010). Widely used to judge concept categorization algorithms, the MNIST dataset is divided into 60,000 training representations and 10,000 test images, with about equal likeness across all ten classes (digits 0–9). The digits were normalized to fit inside a 20×20-pel box while maintaining their facet percentage, and therefore focused within a 28×28 field utilizing center-of-bulk predictions. The dataset was built from NIST's Special Database I and Special Database III, which involve digits composed by grades 9–12 juniors and Census Bureau operators (Deng 2012). A notable feature of MNIST is that the countenances have been pre-treated, accompanying antagonistic aliasing, bearing grayscale levels that capture variations in stroke breadth. This preprocessing creates the dataset more sensitive, and disputing distinguished to twofold pel likenesses.

Sample images from both datasets are illustrated in Figure 1. The Devanagari dataset exhibits blur and low contrast, while the MNIST dataset also contains blurred edges, both of which can negatively impact classification performance. Therefore, preprocessing steps such as image enhancement are crucial before applying deep learning models for classification. To prepare the data for testing and training, we split the datasets into two subsets. Of the Devanagari dataset, roughly 69,000 samples (75%) were used to train the model, with the remaining 23,000 samples, or 25%, left aside for testing. To make it easier to compare the MNIST dataset's results with those of other published studies, we chose the default split of 60,000 training images and 10,000 test images. A fair evaluation of the models' ability to generalize to novel handwriting patterns was ensured by excluding the test images from the training procedure.

Metrics

We use a variety of well-known benchmark metrics that offer an in-depth understanding of the classification behavior during various training and testing stages to fully assess our model. According to recent work on machine learning evaluation, these metrics are particularly helpful for evaluating generalization, identifying overfitting, and resolving issues with imbalanced data (Naidu *et al.* 2023). True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are the four classifications into which the confusion matrix divides classification results as a simple diagnostic tool. In situations where there is a class imbalance or asymmetric misclassification costs, this matrix is very useful since it provides a clear picture of where and how the model misclas-

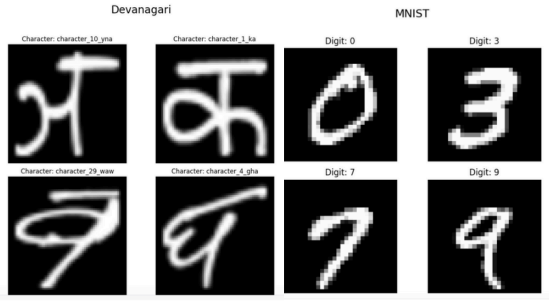


Figure 1 Sample images from datasets

sifies samples. To obtain a more thorough grasp of the model's efficacy, important performance metrics including, precision, recall, and the F1 score, can be extracted from the confusion matrix (Naidu *et al.* 2023).

In applications like spam filtering and fraud detection, where false positives are undesirable, accuracy is especially important. We tracked accuracy and loss curves as a function of training epochs in addition to static performance scores. Accuracy reflects the proportion of correctly classified instances out of the total number of predictions and is mathematically defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

To further refine the performance evaluation, we utilize precision, which measures the proportion of accurate positive predictions out of all instances predicted as positive. It is mathematically expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

To complement accuracy, we employ recall (or sensitivity), which assesses the model's ability to correctly identify all true positive instances. It is quantified as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

A high recall is particularly crucial in applications such as medical diagnosis, where failing to detect a positive case can have severe consequences. Since precision and recall often trade off against each other, the F1 score offers a harmonic mean of the two metrics, providing a single composite measure to evaluate classification performance:

$$F_1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{4}$$

The F1 score is particularly favored in cases of imbalanced datasets, as it effectively balances the trade-off between false positives and false negatives. Recent literature, such as Yacouby & Axman [10], highlights how extensions of these metrics allow for the assessment of model uncertainty and the quality of probabilistic outputs, making them increasingly important in modern machine learning pipelines. While accuracy is intuitive and easy to interpret, it can be misleading in the presence of class imbalance. In such cases, accuracy may appear high even when the model performs poorly on the minority class. To address this, we also monitored the loss, which quantifies the model's prediction error. Effective learning is usually indicated by a significant drop in loss values over training epochs. During backpropagation, model weights are updated using the loss function, such as cross-entropy loss in classification tasks, and loss curves offer a visual representation of the optimization procedure. Overfitting is frequently indicated by a large divergence between the training and validation loss curves, whereas underfitting or less-than-ideal learning rates may be indicated by plateauing curves (Naidu *et al.* 2023). We have plotted the Receiver Operating Characteristic (ROC) curve, which shows the True Positive Rate (TPR) versus the False Positive Rate (FPR) across different decision thresholds, to further evaluate the robustness of classifiers. The equations for these metrics are as follows:

$$True\ Positive\ Rate(TPR\ or\ Recall)(y\ axis) = \frac{TP}{TP + FN} \tag{5}$$

$$False\ Positive\ Rate(FPR)(x\ axis) = \frac{FP}{FP + TN} \tag{6}$$

A scalar measure that assesses the model's capacity for class distinction is the Area Under the Curve (AUC) of the ROC. AUC values vary from 0.5, which denotes random guessing, to 1.0, which denotes perfect classification. Because it considers all potential decision thresholds and offers a thorough summary of performance across various operating points, the ROC-AUC score is especially useful for assessing models trained on imbalanced datasets. These metrics and visualizations allowed us to evaluate our handwritten character recognition models in a comprehensive and multifaceted way. They ensured a solid evaluation of performance by successfully capturing the model's generalization ability as well as classification accuracy (Naidu *et al.* 2023; Yacouby and Axman 2020).

METHODOLOGY

Figure 2 shows the flow diagram for the suggested implementation. We used the Fuzzy added Power-Law Transformation to improve the image quality. For comparative analysis, we also used ABR-based techniques and GLFD (Grunwald-Letnikov Fractional Derivative). Every enhancement method was used consistently throughout the dataset, and distinct models were trained and assessed with the improved photos associated with each technique. The various stages of the implementation, such as preprocessing, image enhancement, data augmentation, and deep learning models, are all depicted in Figure 2.

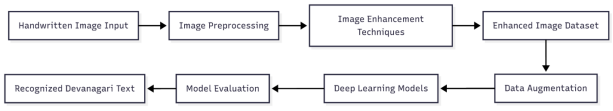


Figure 2 The flow diagram of the proposed implementation

Preprocessing

The first phase of the proposed implementation involves preprocessing. Before feeding the images into the models, a consistent

preprocessing step was applied to all character images. Each image was resized to a fixed resolution to ensure uniformity. For the CNN model, which requires input images of 64×64 pixels, all images were resized to this dimension. This resizing standardizes the aspect ratio and size of the characters, minimizing the impact of variations in the original image sizes on the learning process.

For the MobileNetV2 model, which expects larger color images, the images were upsampled to 224×224 pixels and converted from grayscale to RGB format. This transformation is detailed further in the preprocessing section. After resizing, the images remained grayscale, with pixel intensities ranging from 0 (black) to 255 (white). Binarization was not performed; instead, we retained the grayscale intensity values because subtle grayscale variations can contain valuable information about stroke quality.

Image Enhancement Techniques

Image enhancement plays a crucial role in the preprocessing workflow of this study. We have proposed an enhancement method using Fuzzy added Power-Law Transformation for the input images. Additionally, we implemented GLFD (Grunwald-Letnikov Fractional Derivative) and ABR-based methods with fuzzy to evaluate their impact on recognition. Each method has a distinct function and operation:

Power-Law + Fuzzy (Gamma) Transformation: To enhance the brightness and balance of images, we applied a power-law (gamma) transformation. This transformation is defined as:

$$I_{power\_law} = cI^{\alpha} \tag{7}$$

In equation (7), I represent the pixel intensity in the original grayscale image, α is the gamma exponent, and c is a scaling constant. We typically choose the value of c such that the maximum intensity value of 255 remains unchanged, ensuring that the image's dynamic range is preserved.

The value of α plays a key role in how brightness is redistributed:

- When $\alpha < 1$ the image gets brighter, which helps to bring out darker areas and makes faint strokes more noticeable.
- When $\alpha > 1$ the image darkens, reducing the emphasis on brighter areas and possibly losing some detail in the lighter parts.

As a result of this transformation, the images exhibit generally higher contrast. Backgrounds tend to approach pure white, while text or ink strokes become darker and more distinct. This enhanced visual clarity is expected to assist the neural network by making it easier to differentiate between various character shapes. After applying the power-law transformation, we employed a fuzzy contrast enhancement technique to further emphasize the differences between the foreground and background in grayscale images. This method uses a smooth, non-linear transformation that enhances variations in pixel intensities while preserving structural details. Starting with a grayscale input image, we first normalize the pixel values to fit within the range of [0, 1] by dividing each value by 255:

$$I_{norm} = \frac{I_{power\_law}}{255} \tag{8}$$

Next, we apply the fuzzy transformation, which is defined as:

$$I_{fuzzy} = \frac{I_{norm}^{\gamma}}{\left(I_{norm}^{\gamma} + (I - I_{norm})^{\gamma}\right)} \tag{9}$$

In this equation, γ is a self-intensity factor and is a customizable parameter that adjusts the strength of the contrast enhancement. This setup boosts the contrast around mid-gray levels while ensuring a smooth transition at the extremes. When you increase the value of γ, you get a sharper distinction between light and dark areas. Finally, we rescale the transformed image back to the standard 8-bit range and convert it to an unsigned 8-bit integer format for further processing.

$$I_{enhanced} = I_{fuzzy} \times 255 \tag{10}$$

This fuzzy contrast enhancement technique produces images with sharper feature boundaries, which can significantly improve the performance of tasks such as segmentation or character recognition.

Grunwald–Letnikov Fractional Differentiation (GLFD): Our preprocessing workflow included GLFD (Grunwald-Letnikov Fractional Derivative)-based filtering to enhance image details. To improve intensity variations (like edges and gradients), the GLFD algorithm calculates a fractional derivative of the image intensity function. This is more noticeable than no differentiation, which results in a blurry image, but less aggressive than a full first derivative, which may introduce excessive noise. We successfully sharpened edges in each image by applying a fractional differentiation filter, which made the transitions from strokes to background more defined and steeper. For characters with thin or intricate strokes, this is especially helpful because it keeps them from blending into the background when the image is slightly blurry.

Because it captures fine details without over-accentuating noise, which is a common problem with conventional high-pass filters, the GLFD method is sophisticated. Based on suggestions from the literature, we chose a fractional order (a value between 0 and 1) for the derivative when applying this technique to achieve a noticeable sharpening. GLFD preprocessing produces a sharper image than the original, with improved internal details like holes or loops in the character and more distinct character contours. We expected this to help recognition models better identify the most important structural components of each character class.

For calculating the fractional derivative, Grunwald–Letnikov (GL) formulation is utilized. It is given by:

$$D^{\mu} f(x) = lim_{\epsilon \to 0} \frac{1}{\tau(1-\mu)} \sum_{k=0}^{N} (-1)^k \binom{\mu}{k} f(x - k\epsilon) \tag{11}$$

Where μ s the fractional order of the derivative,$f(x)$is the image intensity at position x,$\epsilon$is a small step size (typically set to 1 pixel), and τ is the Gamma function.$\binom{\mu}{k}$ is the generalized binomial coefficient, defined by:

$$\binom{\mu}{k} = \frac{\tau(\mu+1)}{\tau(k+1)\tau(\mu-k+1)} \tag{12}$$

This structure allows intensity values from earlier times to be weighted, with their influence decreasing gradually based on the fractional order. This well-balanced sensitivity enables richer feature extraction from noisy or low-contrast images by effectively distinguishing between edge saliency and smoothness (Acharya and Gyawali 2015).

ABR Enhancement Approach: The other enhancement approach we employed was ABR (Adaptive Block-based Rescaling) enhancement. To address challenges posed by large physical image sizes, we simplified the problem by assuming that, within each image

block (for example, an 8×8 pixel area), lighting conditions and contrast are equally non-uniform. This assumption enabled us to apply adaptive corrections within blocks: darker or lower-contrast blocks compared to the overall image were brightened or had their contrast stretched, while properly exposed blocks remained unchanged. This adaptive processing is highly effective for images where character parts appear lighter due to uneven or soft scan strokes. The ABR enhancement process involves several steps: fuzzy enhancement, fractional derivatives, and the use of the ABR fractal-fractional gradient mask. Enhancement is performed using a generalized fractional derivative strategy, leveraging the Mittag-Leffler function for calculating fractional derivatives. This approach preserves subtle details without exaggerating noise.

ABR fractal-fractional derivative is given as:

$$F_{ABRD_{\alpha,\beta}}[I(t)] = B(\alpha)\frac{1}{(1-\alpha)^{\beta}}\frac{d^{\beta}}{dt^{\beta}}\int_0^t I(s)E_{\alpha}\left(-\frac{\alpha}{1-\alpha}(t-s)^{\alpha}\right)ds \quad (13)$$

Where α and $]\beta$ represent the fractional order and fractal dimension, respectively, I is the image intensity function at the pixel level.

The Mittag-Leffler function E_{α} is defined as:

$$E_{\alpha}(z) = \sum_{k=0}^{\infty}\frac{z^k}{\tau(\alpha k+1)} \quad (14)$$

Here,z is a complex variable, and τ is the Gamma function, which generalizes the factorial function. This function generalizes the exponential function and is widely applicable in fractional calculus. It also helps in addressing memory effects and self-similarity properties of images (Ramesh Babu *et al.* 2024). Among all the strategies we explored, the most effective combination was Fuzzy enhancement followed by Power-Law transformation, which consistently yielded the best recognition accuracy. This step-by-step approach leverages both local adaptivity and global contrast normalization, resulting in clearer character representations and a significant improvement in model performance.

Data Augmentation

In addition to resizing, we performed data augmentation on the training set. Data augmentation involves generating modified versions of the training images to artificially expand the dataset and introduce more variation. By applying random transformations such as rotations, shifts, shearing, and zooming, we helped the model generalize better to new images. During each training epoch, augmented samples were created dynamically: random small rotations (up to ±10 degrees), minor horizontal and vertical translations (up to 10% of the image size), slight shearing (by a factor of 0.1), and moderate zooming in or out (up to 10%). The robustness of the model is increased by these transformations, which mimic natural variations like characters appearing at different positions or orientations on the paper. A solid basis for developing a successful character recognition model is provided by the combination of augmentation techniques and a varied dataset. We utilized data augmentation during training to enhance model generalization and mitigate overfitting to the training data. Random transformations were applied to each training image with a certain probability in every epoch. The specific details of the data augmentation are as follows:

- Rotation: Random rotations of up to ±10 degrees to simulate characters that may be slanted or rotated during scanning.
- Translation: Small translations of up to 10% of the image size in either direction, enabling the model to recognize characters that are not centrally aligned.
- Shear: A shear transformation (tilting the image) by a factor of 0.1, which mimics changes in writing angle.
- Zoom: Random zooming in or out by up to 10%, allowing the model to handle minor size variations in handwriting.

Deep Learning Models

We evaluated both our custom CNN and MobileNetV2 models on datasets enhanced using each method individually, as well as on the original (unenhanced) data to establish a robust baseline for comparison. We also explored hybrid enhancement techniques, particularly combining Fuzzy contrast enhancement with Power-Law transformation, which yielded the most consistent performance improvements. To ensure fairness across all experiments, we maintained consistent training conditions, such as data augmentation, learning rates, and the number of epochs. Importantly, we applied augmentation after the enhancement step to preserve the quality of the improved images. This systematic approach allowed us to isolate the impact of each enhancement technique on model performance. For example, we observed a notable increase in accuracy for the CNN when using GLFD-based enhancement, while MobileNetV2 showed only minor improvements. This suggests that MobileNetV2's deeper architecture is already effective at capturing fine details, making it less reliant on additional enhancements. These comparisons are essential for determining which enhancement methods work best and for identifying their compatibility with specific models. For our handwritten character recognition project, we examined two deep learning models: a custom Convolutional Neural Network (CNN) and the MobileNetV2 architecture, which we adapted using transfer learning.

Custom CNN Architecture: We designed our CNN specifically to process 64×64 grayscale images. The architecture consists of three convolutional layers followed by two fully connected (dense) layers. Each convolutional block employs 3×3 kernels, ReLU activations, and 2×2 max pooling. The dense layers integrate the extracted features and perform the classification task. To mitigate overfitting, we incorporated a dropout layer, and the final softmax layer outputs probabilities across the 36 character classes. This model is lightweight and easy to train, making it ideal for benchmarking and experimenting with image enhancements. The customized CNN architecture is illustrated in Figure 3 and detailed in Table 1.

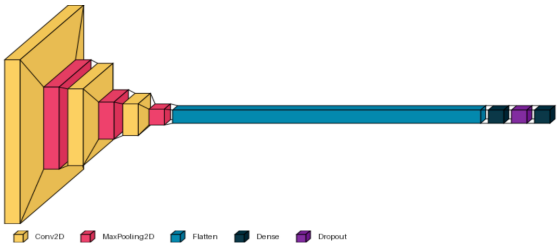


Figure 3 The custom CNN architecture

MobileNetV2: MobileNetV2 is a computationally efficient architecture designed for mobile and embedded systems, offering significant reductions in computational cost without compromising accuracy. We adapted MobileNetV2, a state-of-the-art CNN architecture, for our model. It has been pre-trained on ImageNet, a

Table 1 The custom CNN layered parameters

Layer	Filters / Units	Kernel Size	Activation	Other
Conv2D	32	3×3	ReLU	-
MaxPooling2D	-	2×2	-	-
Conv2D	32	3×3	ReLU	-
MaxPooling2D	-	2×2	-	-
Conv2D	64	3×3	ReLU	-
MaxPooling2D	-	2×2	-	-
Flatten	-	-	-	-
Dense	128	-	ReLU	Dropout (0.5)
Dense (Output)	36	-	-	-

large dataset of labeled images, which enables it to leverage features learned from a wide range of natural images, such as edges, textures, and object parts. This pre-training allows MobileNetV2 to generalize well to new tasks, like handwritten character recognition, especially when working with smaller datasets, using transfer learning.

The MobileNetV2 architecture is detailed in Table 2. Since MobileNetV2 is designed to accept input images of size 224×224 with three color channels (RGB), we modified our grayscale input images accordingly. Specifically, we resized the handwritten grayscale images from 64×64 to 224×224 to match the required input size. To accommodate the three-channel requirement, we duplicated the grayscale channel across all three RGB channels, effectively converting the images into pseudo-RGB format. One of the key innovations in MobileNetV2 is the use of inverted residual bottleneck blocks. These blocks are designed to extract features efficiently while maintaining a lightweight structure. Unlike standard residual blocks, inverted residuals employ depth-wise separable convolutions, which significantly reduce both parameters and computational cost. To accomplish this, a lightweight 1×1 convolution first expands the number of channels; a depth-wise separable convolution then independently convolves each channel; and lastly, another 1×1 convolution shrinks the channels to their initial size. By reducing computations while maintaining high expressiveness, this method allows for rapid and effective processing without sacrificing model accuracy.

The use of ReLU6 activations rather than the conventional ReLU is another noteworthy aspect of MobileNetV2. ReLU6 improves performance and model stability by capping activation values at a maximum of 6, especially in mobile environments where computational efficiency is crucial. Skip connections are another feature of MobileNetV2 that enables the model to learn residual mappings. By avoiding problems like vanishing gradients, these skip connections improve gradient flow during training. Consequently, the model can efficiently learn complex features without experiencing a decline in performance.

We modified MobileNetV2 by adding a unique classification head to the model to tailor it for our handwritten character recognition task. There are two primary parts to the classification head: GAP or global average pooling: By calculating the average value for each channel across the entire map, this layer shrinks the feature map's spatial dimensions (height and width). In this step, the representation is made simpler, fewer parameters are used, and overfitting is avoided without sacrificing important classification-related data. Dense Layer Activated by Softmax: The probabilities for each of the 36 classes in our case, the 36 different Devanagari characters are produced by this layer. Interpretable class probabilities are provided by the Softmax activation, which guarantees that the output values add up to 1.

Following our dataset training, we adjusted the weights in the deeper layers at a low learning rate to refine the model. The model can be fine-tuned to accommodate the unique characteristics of handwritten characters, such as shapes and stroke patterns, without changing the previous layers that capture more general features like edges and textures. By utilizing ImageNet's strong pre-trained weights, this method helps prevent overfitting, particularly when dealing with small datasets. Our handwritten character recognition task was completed by MobileNetV2 with less overfitting and less computational overhead, thanks to its effective architecture, pre-trained weights, and fine-tuning.

Model Environmental Setup and Evaluation

Preprocessing: Before feeding the original or enhanced images into the neural network models, we performed some preprocessing steps to ensure the data was in an optimal format. Here's a breakdown of the key preprocessing techniques: Resizing:

- For the CNN model, all images were resized to 64×64 pixels in grayscale.
- For MobileNetV2, the images were upscaled to 224×224 pixels. To make the grayscale images compatible with MobileNetV2, which expects three-channel (RGB) input, we duplicated the single grayscale channel across all three channels, creating a pseudo-RGB image.
- Resizing ensures that all input images have a consistent size, maintaining the aspect ratio. Since the original images are mostly square and centered, scaling them to a uniform size did not significantly distort the characters. Standardizing the input size also reduces any resolution-related differences that could impact model performance.

Data Augmentation:

- Data augmentation was applied only to the training images to increase variability and improve model robustness. Examples of augmentations include slight rotations (both left and right),

Table 2 MobileNet V2 Architecture

Layer Type	Input Size	Output Size	Kernel Size	Stride	Expansion Factor
Initial Convolution	224×224×3	112×112×32	3×3	2	-
Inverted Residual Block	112×112×32	112×112×16	3×3	1	1
Inverted Residual Block x2	112×112×16	56×56×24	3×3	2	6
Inverted Residual Block x3	56×56×24	28×28×32	3×3	2	6
Inverted Residual Block x4	28×28×32	14×14×64	3×3	2	6
Inverted Residual Block x3	14×14×64	14×14×96	3×3	1	6
Inverted Residual Block x3	14×14×96	7×7×160	3×3	2	6
Inverted Residual Block x1	7×7×160	7×7×320	3×3	1	6
Final Convolution	7×7×320	7×7×1280	1×1	1	-
Global Average Pooling	7×7×1280	1×1×1280	-	-	-
Fully Connected	1×1×1280	1×1×1000	-	-	-

shifts, and other transformations that simulate natural variations in handwriting. These transformations create multiple versions of the same character, helping the model learn more resilient features.
- It is important to note that no augmentation was applied to the test images. The test set remained untouched and static to ensure accurate performance measurement. By expanding the training set through augmentation, we effectively increased its diversity, making the models more adaptable to real-world variations.

Normalization:

- After resizing (and applying augmentations during training), we normalized the pixel intensity values. All pixel values, which originally ranged from 0 to 255 in 8-bit grayscale, were divided by 255 to scale them to floating-point numbers between 0.0 and 1.0.
- This normalization ensures that all input features (pixel intensities) are on the same scale, leading to more stable and efficient training for the neural network. It also prevents large numerical values in raw pixel data from biasing the updates of the network weights.
- In this case, we did not perform mean subtraction or standard deviation scaling because the grayscale data was already quite uniform, and simple division by 255 was sufficient to normalize the inputs effectively.

By carefully preprocessing the images in this manner, we ensured that the data was well-prepared for both the CNN and MobileNetV2 models, enabling them to learn effectively and generalize well to new handwritten characters.

Model Training: We trained both the CNN and MobileNetV2 models under two scenarios: using original images and enhanced images, following a supervised learning approach. Below is a summary of the training configuration and hyperparameters for the custom CNN model:

- Optimizer: We used the Adam optimizer with a learning rate of 0.001. Adam is known for its adaptive learning rate, which typically leads to faster and more stable convergence compared to standard stochastic gradient descent, making it well-suited for our application.
- Loss Function: We employed Categorical Cross-Entropy loss, which is ideal for multi-class classification problems with one-hot encoded labels for our 36-character classes. This loss function measures how far the predicted class probabilities are from the true class (ground truth), guiding the model toward better performance.
- Metrics: During training, we closely monitored accuracy as our primary metric to assess how well the model was fitting the data. Accuracy represents the ratio of correctly classified characters.
- Batch Size: We used mini batches of 32 images during training. This batch size strikes a balance between reducing gradient noise and managing memory requirements. It is small enough to introduce some randomness (which helps the model avoid local minima) but large enough to ensure stability while training on our hardware.
- Epochs: The CNN was trained for a maximum of 30 epochs. We also implemented early stopping if the validation accuracy did not improve for a certain number of consecutive epochs, training would halt to prevent overfitting and unnecessary computation. In practice, the CNN often reached its best performance well before the 30-epoch threshold, due to the dataset size and the use of augmentation.

The configuration and hyperparameters for the MobileNetV2 model were as follows:

- Optimizer: For MobileNetV2, we also used the Adam optimizer, but with a lower learning rate of 0.0001. Since MobileNetV2 was pre-trained, a lower learning rate helped prevent drastic updates that could disrupt the pre-trained features. This subtle update mechanism allowed the model to fine-tune effectively.
- Regularization: Regularization was crucial for improving generalization in both models. For the CNN, we applied dropout in the dense layer, randomly skipping neurons during each training update. This technique prevents the network from over-relying on specific features or feature co-adaptation. MobileNetV2, on the other hand, benefits from its architecture, which includes batch normalization applied during pre-training.
- Early Stopping: To avoid overfitting, we controlled the training duration using early stopping. Additionally, data augmentation served as a regularization method by providing the model with diverse training samples.

During training, we evaluated the model's performance on a validation set after every epoch. This validation set was a subset of the training data reserved specifically for monitoring progress. We saved the model weights when the validation accuracy peaked at any epoch, which we later used to measure performance on the test set and compute final metrics. We repeated this training process for each combination of model and enhancement technique, as well as for a baseline setup without any enhancements. This systematic approach ensured comprehensive evaluation of all configurations.

Model Evaluation: Model performance comparison across different image conditions is a key component of our experimental setup. Specifically, we will compare the performance of both models (CNN and MobileNetV2) on original raw images versus enhanced images processed using three enhancement methods: Fuzzy-Power-Law, GLFD, and ABR-Fuzzy. The original images serve as the baseline for reference. If an enhancement method is effective, the model's accuracy and other performance metrics on the test set should improve compared to the baseline. Based on preliminary observations, Fuzzy-Power-Law, GLFD, and ABR-Fuzzy enhancements have shown promise in increasing recognition accuracy for each model. It remains to be seen whether each model has a unique preference for a specific enhancement method or if one approach consistently outperforms the others.

To present our findings, we plan to provide both quantitative results and qualitative visualizations clearly and understandably. Quantitative results will be condensed using charts. For example, a bar chart might compare CNN's test accuracy across four conditions: No Enhancement, Power-Law, GLFD, and ABR-Fuzzy side by side. Similarly, we will create comparable visualizations for MobileNetV2. These charts will immediately highlight which enhancement method provided the largest improvement for each model.

Qualitative visualizations will complement the quantitative results by illustrating the advantages of individual enhancement techniques. We will present before-and-after images to demonstrate the impact of each enhancement method. For instance, a particularly challenging sample character image can be shown in its raw state alongside versions enhanced using Fuzzy-Power-Law, GLFD, and ABR-Fuzzy. Such comparisons might reveal that ABR-Fuzzy reduces background noise, GLFD emphasizes edges more effectively, and Power-Law makes the image significantly brighter and clearer. These visual representations provide insight into the correlation between image processing and model performance: if an enhancement technique makes a previously illegible character legible, it is reasonable to expect improved model performance on that image.

RESULTS AND DISCUSSION

The experiment was conducted on a computer system equipped with a Python 3.10.12 environment, utilizing TensorFlow 2.15.0, Keras 2.15.0, OpenCV 4.8.0, NumPy 1.24.3, SciPy 1.11.3, Matplotlib 3.7.1, and scikit-learn 1.2.2. GPU acceleration was leveraged to expedite model training and inference, significantly reducing computation time. All preprocessing algorithms and deep learning models were executed within this standardized environment to ensure reproducibility and consistency across different experimental runs. Initial development and testing of the code were performed on an M1 chip. The enhancement techniques were developed and validated using various datasets to assess their performance stability. The MNIST dataset was used as a secondary dataset for further validation in this study, which mainly used the Devanagari script dataset. The findings from these datasets are shown in the ensuing subsections.

Results of Devanagari script

This section presents the CNN and MobileNetV2 models' classification performance after they have been trained and assessed on the original and improved datasets. We methodically implemented three techniques: Power-Law + Fuzzy, GLFD + Fuzzy, and ABR + Fuzzy to evaluate how image enhancement methods affected classification performance. Every result was contrasted with a baseline in which no improvement was made. Table 4 summarizes the overall accuracy of each of these approaches.

Baseline (without enhancement): As shown in the results of Table 4, the CNN model outperformed the MobileNetV2 model on the original dataset without any image enhancements. Despite MobileNetV2's pre-trained, deeper architecture, the CNN achieved an accuracy of 92.12%, surpassing MobileNetV2's baseline accuracy of 90.03%. This outcome is somewhat unexpected, as it suggests that the CNN's simpler architecture may have been better suited to the specific characteristics of handwritten characters. These results provide a crucial baseline for evaluating the improvements observed after applying our image processing techniques.

Power-Law + Fuzzy: The Power-Law + Fuzzy enhancement technique yielded the highest accuracy gains, particularly for the CNN model. To determine the optimal level of contrast adjustment, we tested various values of the power-law parameter α. The results with the CNN are summarized in Table 3. These outcomes indicate an improvement of approximately 6% over the CNN baseline. At $\alpha = 0.6$, the best accuracy of 98.02% was achieved, suggesting that moderate contrast enhancement is most beneficial. Values closer to one yielded diminishing return, whereas values that were too low led to excessive brightening. This improvement highlights how effective preprocessing can compensate for simpler model architecture, enabling the CNN to surpass its baseline and even approach MobileNetV2's baseline accuracy.

Power-Law + Fuzzy preprocessing also enhanced the performance of MobileNetV2, although the improvement was less pronounced, yielding an accuracy of 92.86%. The gain was noticeable but relatively modest because MobileNetV2 is a more robust and pre-trained model that naturally handles a wide range of variations. Nonetheless, this result validates the overall effectiveness of the Power-Law enhancement technique.

Table 3 The Power-Law + Fuzzy enhancement performance comparison over α on Devanagari script

α Value	Accuracy (%)
0.5	97.82
0.6	98.02 (highest)
0.7	97.94
0.8	97.78
0.9	97.82

GLFD + Fuzzy: From the results in Table 4, the GLFD + Fuzzy approach significantly improved recognition accuracy, particularly for the CNN model. The test accuracy of the CNN increased to 95.23%, which is approximately 3% higher than the baseline. This improvement highlights how well GLFD enhances edge features and fine character details, a benefit that is especially advantageous for CNNs, which rely heavily on localized spatial features.

For MobileNetV2, however, the GLFD + Fuzzy approach resulted in a poor accuracy of 76%, likely due to the model's convolutional filters already being highly effective at extracting edge and shape details. This outcome further illustrates the diminishing returns observed when applying enhancements to already high-performing models. While the additional sharpening provided by GLFD + Fuzzy helped slightly, it did not lead to significant improvements. Compared to the baseline, this approach yielded a negative accuracy difference of -14%.

Table 4 Quantitative analysis of Enhancement techniques vs. Deep learning accuracy on the Devanagari dataset

Enhancement Type	CNN Accuracy (%)	Enhancement Type vs. CNN Baseline	MobileNetV2 Accuracy (%)	Enhancement Type vs. MobileNetV2 Baseline
Baseline (Original)	92.12	–	90.03	–
Power-Law + Fuzzy (α=0.6)	98.02	+5.9	92.86	+2.8
GLFD + Fuzzy	95.23	+3.1	76.02	–14.0
ABR + Fuzzy	92.80	+0.6	87.00	–3.03

ABR + Fuzzy: The CNN's accuracy increased slightly to 92.80% with the ABR + Fuzzy enhancement technique. While this represents an improvement over the raw input, it is less effective compared to the results achieved with Power-Law or GLFD. This suggests that ABR + Fuzzy may not always enhance all image types, despite its ability to remove localized noise and strengthen weak strokes. It is possible that some artifacts introduced during the enhancement process offset its advantages, particularly in cleaner samples. When using ABR + Fuzzy preprocessing, MobileNetV2 occasionally showed minimal changes or even slight degradation,

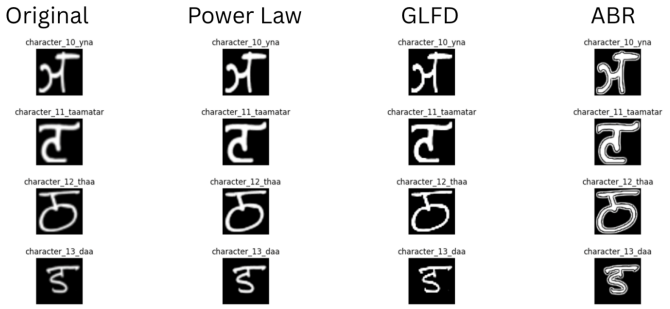


Figure 4 Qualitative results of proposed image enhancement with Devanagari images

resulting in an accuracy of 87%, which is 3.03% lower than the baseline. Over-processing or the introduction of unwanted artifacts may have contributed to this decline, potentially confusing the finely tuned model.

Qualitative results of the proposed image enhancement techniques on original images are presented in Figure 4. Based on the findings from Table 4 and Figure 4, Power-Law + Fuzzy emerges as the most significant improvement overall. It substantially boosts CNN accuracy and narrows the performance gap between CNN and MobileNetV2. GLFD + Fuzzy also offer important improvements, particularly for characters that rely heavily on edges. Although less consistent across all scenarios, ABR + Fuzzy demonstrates certain benefits when dealing with noisy or degraded inputs.

Therefore, Power-Law + Fuzzy emerge as a well-suited enhancement technique for the Devanagari script. The other qualitative metrics are presented in Table 5. The final CNN model's classification report shows a macro-averaged precision of 98.12%, recall of 98.09%, and an F1-score of 98.09% across the 36-character classes, indicating good overall generalization. Several classes, such as character_13_daa, character_22_pha, and character_36_gya, achieved perfect precision and recall, demonstrating that the model effectively generalized the distinct structural features of these characters. The results in Table 5 were computed using the confusion matrix obtained from the Power-Law + Fuzzy enhancement. The confusion matrix is shown in Figure 5, which reveals that most misclassifications occur between structurally similar characters, particularly in noisy or lightly stroked samples.

Table 5 Quantitative analysis of Power-Law + Fuzzy enhancement on Devanagari script

Metric	Value
Accuracy	98.02%
Precision	98.14%
Recall	98.09%
F1-Score	98.09%

The ROC curves for each class, displayed in Figure 6, show near-perfect separation for most classes, indicating strong confidence calibration. The training and validation loss curves demonstrate smooth convergence with minimal overfitting, reflecting the effectiveness of our preprocessing steps in stabilizing the learning process.

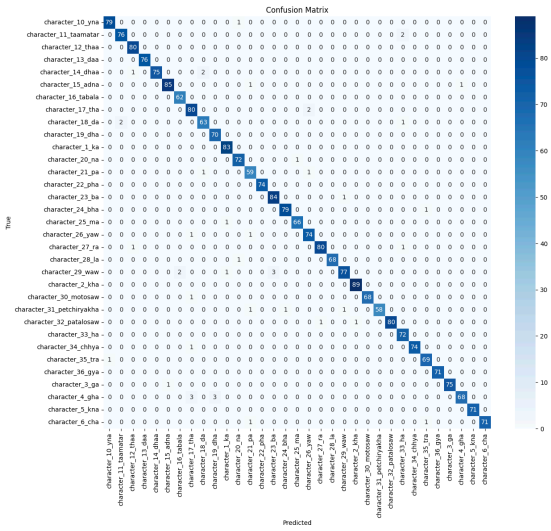


Figure 5 The confusion matrix of Power-Law + Fuzzy enhancement on Devanagari script

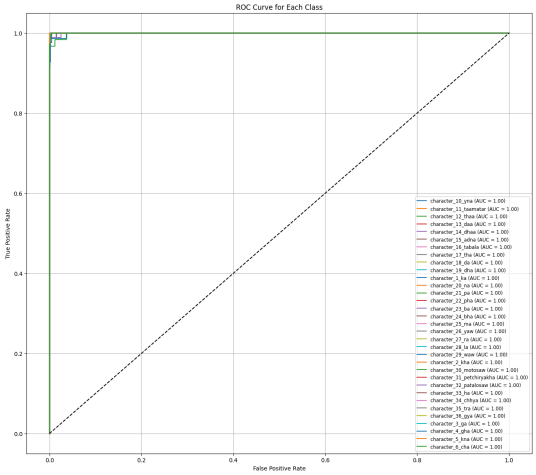


Figure 6 The ROC curve of Power-Law + Fuzzy enhancement on Devanagari script

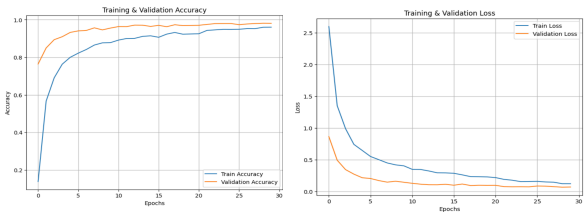


Figure 7 The accuracy and loss curve vs. epoch on Devanagari script using Power-Law + Fuzzy enhancement

Most importantly, these findings show that, with appropriately augmented input, a lightweight CNN can match or even surpass the performance of a heavier model. This has significant practical implications in environments with limited computational resources, where preprocessing at the input level can serve as an extremely effective and cost-efficient method for enhancing recognition accuracy.

Table 6 Quantitative analysis of Power-Law + Fuzzy enhancement on Devanagari script

Research	Method	Accuracy (%)
Narang et al. Narang *et al.* (2021)	Custom CNN	93.73
Sarangi et al. Sarangi *et al.* (2020)	SVM	95.65
Hasan et al. Hasan *et al.* (2023)	Two-stage VGG16	96.55
Proposed Work	Power-Law + Fuzzy and CNN	98.02

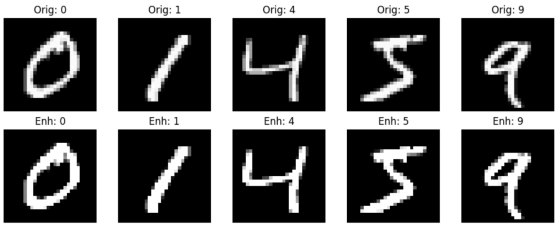


Figure 8 The qualitative results of Power-Law + Fuzzy enhancement on the MINST dataset

Table 6 summarizes recent literature and its findings related to the Devanagari script. Narang et al. developed a specialized model called DeepNetDevanagari for recognizing ancient Devanagari characters, achieving an accuracy of 93.73% (Narang *et al.* 2021). Their work addressed specific challenges associated with degraded ancient documents but did not incorporate any preprocessing techniques to improve image quality before classification. In contrast, Sarangi et al. took a different approach by using feature selection methods combined with SVM classifiers instead of deep learning (Sarangi *et al.* 2020). Their solution achieved an accuracy of 95.65% on the dataset, demonstrating that effective feature engineering can enable older machine learning methods to yield competitive results, though still not matching the performance of our enhancement-based method. Hasan et al. achieved an accuracy of 96.55% using a two-stage VGG16 network with adaptive gradient methods (Hasan *et al.* 2023). Their technique leveraged transfer learning and fine-tuning to reduce the number of trainable parameters, making it computationally efficient, although it lacked the preprocessing advantages offered by our method.

The proposed Power-Law + Fuzzy enhancement method demonstrates superior performance compared to other recent approaches in Devanagari character recognition. The key innovation

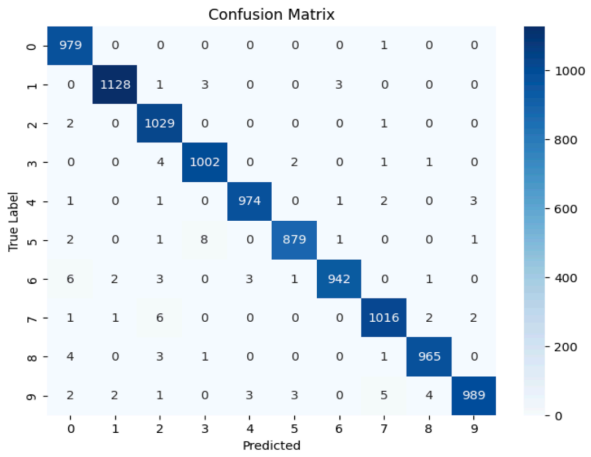


Figure 9 The confusion matrix of Power-Law + Fuzzy enhancement on the MINST dataset

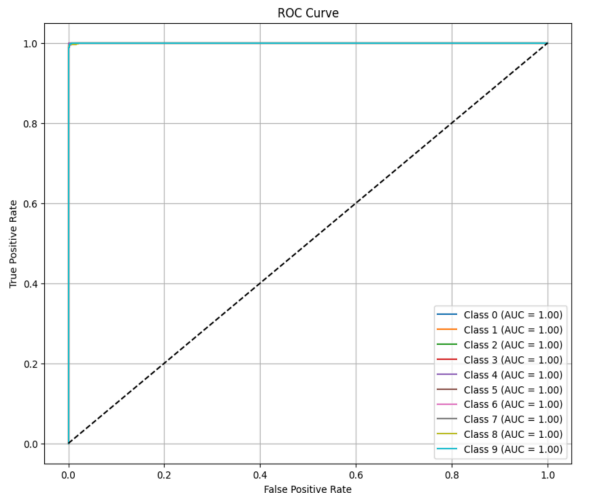


Figure 10 The ROC curve of Power-Law + Fuzzy enhancement on the MINST dataset

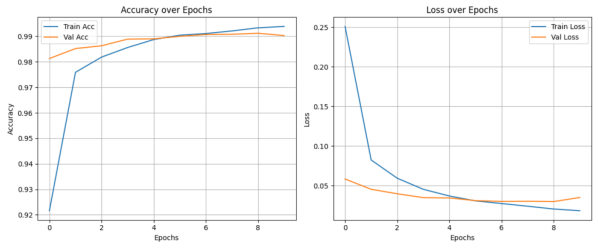


Figure 11 The accuracy and loss curve vs. epoch on MINST dataset using Power-Law + Fuzzy enhancement

of our approach lies in its focus on advanced image preprocessing, rather than relying heavily on complex neural network architectures or intricate feature engineering. A unique contribution of our research is the combination of Power-Law transformation with Fuzzy enhancement, which improves character contrast and edge definition before classification. By prioritizing preprocessing, our method achieves exceptional accuracy (98.02%) using a simple CNN model, making it both computationally efficient and effective when compared to more elaborate methodologies.

Results of the MNIST Dataset: To further evaluate the robustness and general applicability of the proposed Power-Law + Fuzzy enhancement technique integrated within the CNN model, we extended our experiments to include the MNIST dataset. This dataset consists of grayscale images of handwritten digits. Following the preprocessing steps previously applied to the Devanagari dataset, MNIST images were resized to 64×64 pixels to match the CNN input requirements. The enhancement process was carried out in two sequential stages first applying fuzzy enhancement, followed by the adaptive Bernstein Re-weighted (ABR) derivative ensuring a consistent enhancement approach across datasets. Importantly, the CNN architecture remained unaltered to ensure fair and unbiased comparison between the original and enhanced image inputs.

The MNIST dataset used in our experiments comprises 70,000 grayscale digit images (from 0 to 9), categorized into 10 classes. We adopted the standard data split: 60,000 images for training and 10,000 for testing, maintaining a 6:1 train-test ratio. Each digit class is evenly represented, with about 6,000 training images and 1,000 test images per class. This balanced distribution ensures that the model is adequately trained and evaluated across all classes. The standardized nature of MNIST also enables direct benchmarking against existing research.

We trained two CNN models under identical conditions: one using the original MNIST images and the other using the enhanced versions. Both models were trained over 10 epochs using the same data splits and hyperparameter settings. As shown in Figure 8, a visual comparison reveals that the enhanced images exhibit crisper edges and improved contrast, aiding the CNN in capturing class-specific features more effectively.

The final classification results on the MNIST dataset showed that the baseline CNN achieved 99.11% accuracy without enhancement, whereas the model using Power-Law + Fuzzy enhancement achieved a slightly higher accuracy of 99.15%. Although the numerical improvement is small, it demonstrates the consistent benefits of the enhancement technique, even on datasets that are already well-structured and optimized, like MNIST. This modest gain underlines the method's ability to amplify discriminative features, thereby supporting better generalization by the CNN.

To support the previously reported accuracy results, the corresponding confusion matrix is presented in Figure 9. The matrices indicate that the model produced very few misclassifications overall. Notably, the enhanced model demonstrated reduced confusion between visually similar digits, further validating the effectiveness of the enhancement technique.

Figure 10 displays the ROC curve for the Power-Law + Fuzzy enhancement applied to the MNIST dataset. The Receiver Operating Characteristic (ROC) curves for each digit class indicate that the model achieved near-perfect class separation, with area under the curve (AUC) scores reaching or approaching 1.0. This result highlights the effectiveness of the enhancement technique in improving feature discrimination. Figure 11 presents the accuracy and loss curves for the MNIST dataset. Throughout the training

epochs, the model trained on enhanced images consistently outperformed the baseline in both validation accuracy and validation loss, demonstrating more stable and reliable convergence behavior.

Overall, the findings indicate that while the numerical improvements may appear modest, the enhancement methods consistently contribute to increased model robustness and improved convergence behavior. Importantly, the proposed image enhancement approach proves effective not only for complex scripts such as Devanagari but also generalizes well to simpler, widely used datasets like MNIST. These outcomes collectively affirm the efficacy and adaptability of the enhancement technique in the context of handwritten character recognition.

Table 7 Quantitative analysis of Power-Law + Fuzzy enhancement on Devanagari script

Research	Method	Accuracy (%)
Al-Ataby et al. Al-Ataby *et al.* (2021)	Standard SVM	96.30
Kumar et al. Kumar *et al.* (2021)	Custom CNN	97.07
Zhao and Liu Zhao and Liu (2020)	LeNet-5	98.00
Proposed Work	Power-Law + Fuzzy and CNN	99.15

Table 7 presents a comparison between the proposed results and selected recent studies. Al-Ataby et al. explored classical machine learning approaches for handwritten digit recognition, achieving an accuracy of 96.30% using a Support Vector Machine (SVM) model (Al-Ataby *et al.* 2021). Their findings highlighted the limitations of traditional machine learning techniques when applied to the MNIST dataset, especially compared to deep learning methods. Kumar et al. implemented a basic CNN architecture for MNIST digit classification and reached an accuracy of approximately 97.07% (Kumar *et al.* 2021). Their experiments investigated how variations in the number of convolutional layers and filter sizes affected performance. However, their results indicated that even straightforward CNN configurations without specific tuning often fail to surpass the 98% accuracy mark. Zhao and Liu adopted the well-known LeNet-5 deep learning model for MNIST digit recognition, reporting an accuracy of 98.10% (Zhao and Liu 2020). While their study demonstrated the effectiveness of this established architecture, it also introduced added complexity. In contrast, our Power-Law + Fuzzy enhancement approach outperforms these methods, further underscoring the importance of robust preprocessing in boosting classification accuracy, even when applied to well-established CNN models.

The proposed Power-Law + Fuzzy enhancement method achieved a test accuracy of 99.15% on the MNIST dataset, significantly outperforming the alternative approaches discussed. These comparisons highlight the strength of our enhancement technique in improving input data quality through effective image preprocessing. By enhancing discriminative features, the method enables even a relatively simple CNN architecture to attain superior recognition performance.

LIMITATIONS AND FUTURE WORK

While the study demonstrates successful implementation of novel techniques and impressive system performance, it is important to acknowledge several limitations. No single enhancement method consistently produced the best results across all images. For instance, the ABR-Fuzzy approach sometimes diminished performance when applied to already clean images, while an overly aggressive Power-Law adjustment occasionally eliminated important image details. This suggests that our enhancements were tuned to improve the average performance across the dataset, but not necessarily optimized for every individual image. In some cases, certain images may have received little to no benefit from the enhancement process. This indicates that a general, one-size-fits-all approach may not be ideal, and incorporating adaptive enhancement strategies could yield even better results.

During training, particular care was needed to prevent overfitting, especially when using the enhanced dataset with augmented inputs. Feeding the CNN with higher-quality and augmented images increased the risk of the model memorizing the training data, particularly when training was extended for too many epochs. To mitigate this, we implemented dropout layers and early stopping techniques, which ultimately enabled good generalization to the test set. However, the need for such precautions highlights how powerful enhancements can potentially exaggerate training set noise or artifacts, leading to overfitting. Interestingly, overfitting was less problematic in experiments involving the pre-trained MobileNetV2 model, likely due to its prior exposure to broader data and reduced capacity gains from our modifications.

Our experiments were conducted on a specific Devanagari character dataset comprising 92,000 samples. Although sizable and well-curated, this dataset may not capture the full spectrum of handwriting styles or scanning variations. Consequently, the models might have internalized certain biases unique to this dataset. For example, performance might degrade when classifying characters written with different instruments or on rougher paper. Additionally, our enhancement techniques, such as the Power-Law adjustment were calibrated within a fixed parameter range, which may not be optimal for other datasets. Thus, one key limitation is that the proposed system was not evaluated on additional datasets or real-world inputs outside of the current scope. Stronger generalization would require testing across a more diverse range of handwriting conditions and data sources.

Despite these limitations, the current study lays a strong foundation for future improvements. The identified weaknesses point to several promising directions, such as incorporating adaptive enhancement strategies, expanding the number of character classes, and testing on varied dataset sizes. One intriguing possibility is to tailor enhancement techniques based on stroke characteristics, such as thickness or noise level, allowing the system to adaptively select preprocessing methods. For instance, a fuzzy logic controller could be introduced to determine the most suitable enhancement type: Power-Law for dark images, GLFD for overly bright and blurry ones and ABR for noisy inputs. A conditional processing pipeline like this could optimize preprocessing on a per-image basis, potentially leading to more robust classification outcomes.

In this work, MobileNetV2 was chosen as an efficient and compact CNN model. Future research could explore alternative architectures, such as EfficientNet, known for its strong accuracy-to-parameter ratio, or ResNet50, which offers deeper representation capabilities. Additionally, vision transformers might be considered, especially if the research shifts toward phoneme or full-text recognition. Each architecture may respond differently to prepro-

cessing strategies, and certain models could be more resilient to noise or benefit uniquely from specific enhancements. Another worthwhile direction is to extend beyond isolated character recognition to word- and line-level recognition in Devanagari text. Ultimately, our character recognition module could be integrated into a larger OCR pipeline, where image enhancement also supports segmentation tasks.

In conclusion, there remains a significant opportunity for future work to build upon this research. A recurring theme is the synergistic interplay between image preprocessing and deep learning, which has proven beneficial in this study. The results affirm that enhancing input quality through intelligent preprocessing can meaningfully boost recognition performance. Future studies can further explore and refine this synergy, expanding its application to more languages, character sets, and real-world handwritten recognition systems. The foundation laid here demonstrates the potential of such integrated approaches to substantially elevate the accuracy and generalizability of handwriting recognition technologies.

CONCLUSION

Handwritten character recognition is essential in many fields, such as digitizing documents, preserving manuscripts, and developing autonomous text-reading systems. This study focuses on recognizing handwritten Devanagari script, a traditional writing system that originated from the Brahmi script and is widely used for languages like Hindi, Marathi, Nepali, and Sanskrit across South Asia. Because Devanagari is extensively used in government records, educational materials, and cultural texts, accurately identifying these characters is crucial for preserving linguistic heritage, automating Optical Character Recognition (OCR) systems, and enabling multilingual document processing and data entry. To achieve these goals, this research employs deep learning techniques, particularly optimized Convolutional Neural Networks (CNNs), and combines them advanced image processing methods to enhance recognition performance.

This study presents fuzzy logic-based image enhancement techniques, such as Power-Law transformation, Grunwald–Letnikov Fractional Differentiation (GLFD), and Atangana–Baleanu–Riemann (ABR) differentiation, to improve recognition accuracy. By enhancing contrast, sharpening edges, and lowering noise, these techniques aim to improve image features. Two classification models, the lightweight MobileNetV2 and a customized CNN architecture were assessed. A publicly available dataset of handwritten Devanagari characters was used to train and assess both models. Results from experiments show that image preprocessing significantly affects classification performance. When paired with Power-Law + Fuzzy enhancements, the custom CNN produced an astounding accuracy of 98.02%, significantly surpassing MobileNetV2's 92.86%. Overall, the Power-Law + Fuzzy combination produced the best results, even though GLFD and ABR also helped to improve performance.

These findings demonstrate that incorporating well-designed preprocessing methods into even relatively simple CNN architectures can result in high recognition accuracy. The proposed approach is not only effective for complex scripts such as Devanagari, but it also generalizes well, as evidenced by its 99.15% accuracy on the MNIST handwritten digit dataset. This demonstrates the versatility and robustness of the enhancement techniques across diverse datasets. Overall, this framework shows great promise for practical applications in OCR systems for Devanagari and other scripts, providing an efficient and scalable solution for multilingual document analysis and recognition.

Acknowledgments

J. Alzabut expresses his sincere thanks to Prince Sultan University for supporting this research.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Acharya, S. and P. Gyawali, 2015 Devanagari handwritten character dataset. UCI Machine Learning Repository.

Agrawal, K. K. and M. S. Nair, 2019 Fractional order edge detection with gl fractional derivative mask. Signal, Image and Video Processing **13**: 27–34.

Al-Ataby, A., W. Al-Nuaimy, and M. Al-Taee, 2021 Handwritten digit recognition using support vector machine and other traditional machine learning approaches. Engineering Technology Journal **39**: 1131–1141.

Arora, S., L. Malik, S. Goyal, D. Bhattacharjee, M. Nasipuri, *et al.*, 2024 Devanagari character recognition: A comprehensive literature review. IEEE Access .

Atangana, A. and D. Baleanu, 2016 New fractional derivatives with nonlocal and non-singular kernel: Theory and application to heat transfer model. Thermal Science **20**: 763–769.

Das, S. and P. Gupta, 2013 Edge detection using fractional order differentiation and its applications. Procedia Computer Science **17**: 301–308.

Deng, L., 2012 The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**: 141–142.

Garg, V. and K. Singh, 2012 An improved grunwald-letnikov fractional differential mask for image texture enhancement. International Journal of Advanced Computer Science and Applications **3**: 130–135.

Gonzalez, R. C. and R. E. Woods, 2002 *Digital Image Processing*. Prentice Hall, second edition.

Hasan, M. M., M. A. Hossain, A. Y. Srizon, and A. Sayeed, 2023 Devanagari handwritten character recognition using fine-tuned deep convolutional neural network on trivial dataset. Multimedia Tools and Applications **82**: 20671–20686.

Howard, A. G. *et al.*, 2017 Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Jayadevan, R., S. R. Kolhe, P. M. Patil, and U. Pal, 2011 Offline recognition of devanagari script: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **41**: 782–796.

Khaparde, M. S. and V. H. Mankar, 2018 A comprehensive review on ocr for handwritten devanagari script. International Journal of Engineering and Technology **7**: 31–37.

Krizhevsky, A., I. Sutskever, and G. E. Hinton, 2017 Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**: 84–90.

Kumar, R., A. Gaur, K. Bhushan, and M. Singh, 2021 Knowledge extraction in digit recognition using mnist dataset: Evolution in handwriting analysis. International Journal of Applied Engineering Research **16**: 845–852.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner, 1998 Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**: 2278–2324.

LeCun, Y., C. Cortes, and C. J. C. Burges, 2010 Mnist handwritten digit database. Available at http://yann.lecun.com/exdb/mnist.

Liu, C. L., K. Nakashima, H. Sako, and H. Fujisawa, 2003 Handwritten digit recognition: Benchmarking of state-of-the-art techniques. Pattern Recognition **36**: 2271–2285.

Naidu, G., T. Zuva, and E. M. Sibanda, 2023 A review of evaluation metrics in machine learning algorithms. In *Artificial Intelligence Application in Networks and Systems*, edited by R. Silhavy and P. Silhavy, Springer, Cham.

Narang, S. R., M. Kumar, and M. K. Jindal, 2021 Deep net devanagari: A deep learning model for devanagari ancient character recognition. Multimedia Tools and Applications **80**: 20671–20686.

Pal, U. and B. B. Chaudhuri, 2004 Indian script character recognition: A survey. Pattern Recognition **37**: 1887–1899.

Podlubny, I., 1999 *Fractional Differential Equations*. Academic Press.

Ramesh Babu, N., A. S. Joshua, P. Balasubramaniam, and A. Tiwari, 2024 Fuzzy-driven image enhancement via abr-fractal-fractional differentiation. Information Sciences **675**: 120741.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, 2018 Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sarangi, P. K., S. Singh, and C. Singla, 2020 Feature selection for character recognition of handwritten devanagari and odia scripts. International Journal of Engineering Research and Technology **13**: 1974–1982.

Sethi, N. and A. Singh, 2020 Handwritten devanagari character recognition using deep convolutional neural network. Procedia Computer Science **173**: 215–222.

Sharma, S. and D. Kumar, 2021 A fuzzy-based abr image enhancement technique for improved handwritten character recognition. Journal of Ambient Intelligence and Humanized Computing **12**: 10457–10469.

Yacouby and Axman, 2020 Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of Eval4NLP*.

Zhao, H. H. and H. Liu, 2020 Multiple classifiers fusion and cnn feature extraction for handwritten digits recognition. Granular Computing **5**: 411–418.

How to cite this article: Akshara, S., Vinodkumar, A., Sriramakrishnan, P., Saravanan, A., and Alzabut, J. Integrating Fuzzy Logic and Fractional-Order Operators in Convolutional Neural Networks for Handwritten Digits and Characters Recognition. *Chaos and Fractals*, 3(2), 92-107, 2026.

ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.108-116
https://doi.org/10.69882/adba.chf.2026074

Diffusion, Not Chaotic Forcing, Controls Mixing in a Boundary-Coupled Cellular Automaton

Mustafa Kutlu *,1 **and Mehmet Kutlu** α,2

*Department of Systems Engineering, Military Technological College, Muscat, Oman, αClinic for General and Gerontopsychiatry, Oberberg Hospital, Gummersbach, Germany.

ABSTRACT A recurring intuition is that chaotic boundary forcing mixes a medium more thoroughly than regular or random forcing. We test this in a boundary-driven probabilistic cellular automaton, quantifying mixing via Shannon entropy and Moran's I, and find the intuition fails. Experiments demonstrate that neither coupling strength nor forcing character (chaotic Lorenz, white-noise, periodic, or static) significantly affects the final entropy. Instead, a gradient-boosted surrogate ($R^2 = 0.81$) and SHAP attribution identify the internal diffusive update weight as the dominant mixing driver. Exact spectral analysis confirms this mechanism: boundary drives perturb bulk functionals only at order $1/N$, whereas the diffusion weight heavily controls modal damping. While steering this diffusion weight recovers ninety-five per cent of mixing performance, an advective benchmark shows true chaotic velocity fields mix six times faster. This localizes our null result to additive boundary forcing rather than chaotic transport itself. Ultimately, bulk mixing in this system is governed by internal diffusive transport rather than the boundary drive.

KEYWORDS
Chaotic mixing
Cellular automata
Surrogate modelling
Explainable AI

INTRODUCTION

The idea that chaos promotes mixing is one of the most productive in nonlinear science. Since the recognition that even a laminar, deterministic velocity field can stretch and fold material lines exponentially, chaotic advection has provided the theoretical backbone for understanding how fluids and granular media homogenise when stirring is slow and molecular diffusion is weak (Aref 1984; Ottino 1989, 1990; Aref 2002). The appeal of the mechanism lies in its economy: a simple, low-dimensional forcing, applied at a boundary or through a time-dependent perturbation, can in principle drive efficient transport throughout a domain.

It is therefore natural to expect that a chaotic driver, with its broadband spectrum and sensitive dependence on initial conditions, will outperform a periodic or a purely random one when the objective is to mix. This expectation is widely held, and it motivates the recurring use of chaotic signals as actuation in mixing devices and models. A distinction must be drawn at the outset, however, between chaotic *advection*, in which a chaotic velocity field transports material throughout the bulk, and the *additive boundary forcing* studied here, in which a chaotic signal is injected into a single boundary row with no accompanying velocity field. The former is the mechanism of the classical theory; the latter is the weaker and more common engineering surrogate, and whether it inherits any of the mixing virtue of the former is the question we pose. We are careful, therefore, not to conflate a null result for boundary forcing with a claim about chaotic advection, and we include an advective benchmark below precisely to keep the two apart.

Whether that expectation survives contact with a discrete, dissipative medium is less clear. Cellular automata offer a transparent laboratory for the question, because their local update rules make the competing transport channels, self-retention, neighbour averaging, diffusion and external drive, explicit and separately tunable (Wolfram 1983; Chopard and Droz 1998). In such a medium the boundary forcing is only one of several mechanisms redistributing concentration, and there is no guarantee that it dominates the others. A driver that injects structure at a single boundary must compete with the automaton's own relaxation, and if internal transport is efficient the details of the drive may be washed out before they influence the bulk. Distinguishing these possibilities requires more than a single simulation; it requires that the strength of the coupling and the nature of the forcing be varied systematically and that the response be measured with well-defined, reproducible metrics.

Two such metrics recur across the literature on spatial homogenisation. The Shannon entropy of the concentration histogram measures how uniformly mass is distributed over its accessible range and rises as the field mixes (Shannon 1948), while

Manuscript received: 3 February 2026,
Revised: 5 June 2026,
Accepted: 7 July 2026.

[1]mustafa.kutlu@mtc.edu.om (**Corresponding author**)
[2]mehmet.kutlu@klinikum-oberberg.de

Moran's I measures spatial autocorrelation and falls toward zero as neighbouring regions decorrelate (Moran 1950). Used together they separate the amount of mixing from its spatial texture, and their combination provides a compact, quantitative readout of state that is well suited to automated experimentation across many parameter settings.

Establishing which parameters matter is, however, only half of an explanation. Machine-learning surrogates now make it feasible to learn the map from a model's parameters to its emergent behaviour and to interrogate that map for the variables that carry the most weight (Breiman 2001; Friedman 2001; Chen and Guestrin 2016). Attribution methods based on Shapley values give a principled, additive decomposition of a prediction across its inputs (Lundberg and Lee 2017; Lundberg *et al.* 2020), and when their rankings agree with an independent measure such as permutation importance (Altmann *et al.* 2010) the attribution can be trusted as a statement about the surrogate rather than an artefact of the method. A further and more demanding test is whether an identified driver can be exploited: if a variable genuinely controls the outcome, then steering it, for instance within a model-predictive loop (Mayne *et al.* 2000; Camacho and Bordons 2007), should improve that outcome. Explanation and control are distinct claims, and a rigorous study should keep them apart (Ribeiro *et al.* 2016; Doshi-Velez and Kim 2017).

This paper brings these threads together on a single, deliberately simple system. We couple a two-dimensional probabilistic cellular automaton to a chaotic Lorenz driver through one boundary row, with an adjustable coupling amplitude, and we ask what actually determines how well the automaton mixes. We first test the chaos-promotes-mixing intuition directly, by sweeping the coupling strength and by ablating the forcing against noise, periodic and static alternatives. We then build a gradient-boosted surrogate over the automaton's internal update weights, attribute its predictions with SHAP, and validate the attribution against permutation importance. Finally, we ask whether the identified driver confers control leverage by embedding the SHAP weights in a model-predictive controller and comparing it against an unguided baseline. Underpinning the empirical study is an exact spectral analysis of the update operator, which because it is translation-invariant on the torus is diagonalized in closed form; the resulting symbols predict each experimental outcome quantitatively and reduce the null results to a surface-to-volume argument. The analysis returns a clear and somewhat deflationary answer: within this class of boundary-driven automaton, mixing is governed by internal diffusive transport and is essentially indifferent both to the strength of the boundary coupling and to whether that boundary is driven chaotically, randomly, periodically or not at all.

MODEL AND METHODS

Chaotic boundary driver

The forcing signal is drawn from the Lorenz system,

$$\dot{x} = \sigma\,(y - x), \tag{1}$$
$$\dot{y} = x\,(\rho - z) - y, \tag{2}$$
$$\dot{z} = x\,y - \beta\,z, \tag{3}$$

with the classical parameters $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, integrated by an adaptive Runge–Kutta scheme from the initial condition $(0.1, 0, 0)$. To confirm that the driver operates in the chaotic regime we estimated its largest Lyapunov exponent by the standard Benettin two-trajectory procedure with renormalisation, obtaining $\lambda_1 = 0.93$ nats s^{-1} (1.34 bit s^{-1}), in agreement with the canonical Lorenz value and certifying the exponential separation of nearby states that defines deterministic chaos (Lorenz 1963; Strogatz 2015). The $x(t)$ component, rescaled to the interval $[-1, 1]$, supplies the boundary drive fed to the automaton (Figure 1).

Boundary-coupled cellular automaton

The medium is a square 40×40 lattice carrying a scalar concentration $c_{ij} \in [0, 1]$, initialised with the upper half at unity and the lower half at zero to create a sharp, maximally unmixed interface. At each step the field is updated by a convex combination of four local channels,

$$c' = w_0\,c + w_1\,\mathcal{A}c + w_2\,\mathcal{L}c + w_3\,\mathcal{D} + \eta, \tag{4}$$

where $\mathcal{A}c$ is the four-neighbour average, $\mathcal{L}c$ is the discrete Laplacian, and $\mathcal{D}$ injects the boundary drive into the top row with amplitude α, so that $\mathcal{D}_{0j} = \alpha\,\tilde{x}$ with $\tilde{x}$ the rescaled Lorenz sample and $\mathcal{D}_{ij} = 0$ elsewhere. The term η is a small Gaussian noise of standard deviation 0.02, and the result is clipped to $[0, 1]$ after each update. The weights $w = (w_0, w_1, w_2, w_3)$ are non-negative and normalised to sum to unity; they represent, respectively, self-retention, neighbour averaging, diffusion and the coupling to the chaotic drive. Separating the neighbour-average and Laplacian channels is deliberate, since the former redistributes concentration while conserving local mean and the latter drives it down its gradient, and the two need not act with equal strength.

Mixing metrics

Two scalar diagnostics summarise the state of the field. The Shannon entropy

$$H = -\sum_b p_b\,\ln p_b \tag{5}$$

is computed over a twenty-bin histogram of the concentration values, with p_b the normalised bin occupancy; it is bounded above by $\ln 20 \approx 2.996$ nats and increases as mass spreads across its range. The spatial autocorrelation is measured by Moran's I,

$$I = \frac{n}{W}\,\frac{\sum_i \sum_j w_{ij}\,(c_i - \bar{c})(c_j - \bar{c})}{\sum_i (c_i - \bar{c})^2}, \tag{6}$$

with a four-neighbour contiguity weighting; it approaches $+1$ for a clustered, unmixed field and falls toward zero as neighbouring cells decorrelate. Reported values are taken at the final step of each run.

Coupling sweep and forcing ablation

The chaos-promotes-mixing hypothesis was tested in two designed experiments. In the coupling sweep the amplitude α was varied over $\{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ with the update weights held at a fixed reference vector, and each setting was repeated over twenty-five independent random seeds so that the effect of coupling could be read against the natural replicate scatter. In the forcing ablation the coupling amplitude was fixed at $\alpha = 0.4$ and the Lorenz drive was replaced, in turn, by white noise of matched variance, a sinusoid of comparable amplitude, and a static zero drive, with thirty replicates per condition. Differences in final entropy against the Lorenz reference were assessed by Welch's unequal-variance t-test.

Chaotic advection benchmark

To separate additive boundary forcing from genuine advective transport we ran a companion experiment in which the same

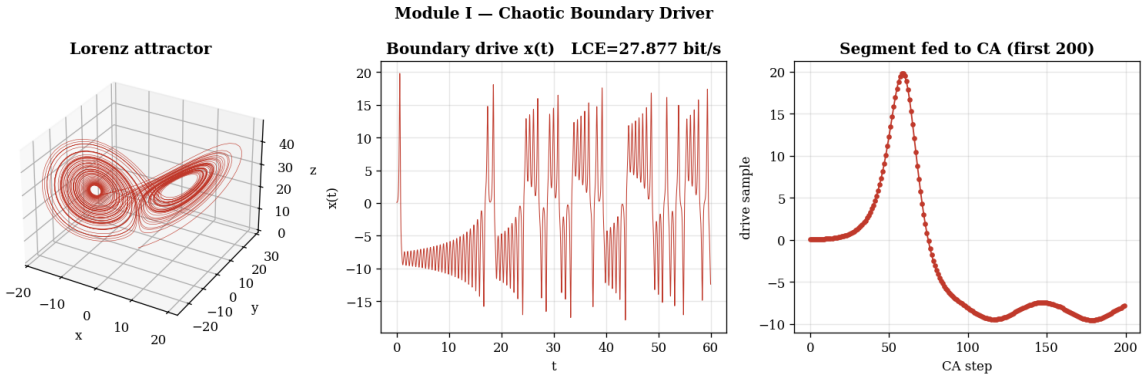


Figure 1 The chaotic boundary driver. Left: the Lorenz attractor. Centre: the $x(t)$ time series, whose largest Lyapunov exponent $\lambda_1 = 0.93$ nats s^{-1} confirms chaotic dynamics. Right: the leading segment of the rescaled drive that is applied to the automaton's top boundary row

Lorenz signal drives a bulk velocity field rather than a boundary source. The scalar was advected by an incompressible sine flow, $u = U\sin(2\pi y + \pi\tilde{x}_t)$ and $v = U\sin(2\pi x + \pi\tilde{z}_t)$, whose phases are modulated by the rescaled Lorenz components, using semi-Lagrangian back-tracing with periodic bilinear interpolation and a small molecular diffusivity $\kappa = 0.02$. Because pure advection is measure-preserving and does not by itself broaden the concentration histogram, mixing was quantified by the decay of the scalar variance, whose half-life $t_{1/2}$ measures how quickly filamentation thins gradients into the diffusive range. The chaotic drive was compared against a periodic flow and against pure diffusion with the advection disabled.

Surrogate model and attribution

To identify the internal determinant of mixing we sampled the four update weights from a symmetric Dirichlet distribution and ran the automaton to completion for each of six hundred weight vectors, recording the final entropy as the target. Six features were formed from the four weights together with two informative ratios, the neighbour-to-self ratio and the diffusion share, the latter defined as $w_2/(w_2 + w_3)$. Six regressors were compared under five-fold cross-validation: an ordinary linear model and degree-two and degree-three polynomial models, included specifically to test whether the weight-to-entropy map is simple enough to dispense with a flexible learner, alongside a random forest, gradient boosting and gradient-boosted trees; the strongest was retained as the surrogate. Two ninety-five per cent prediction intervals were constructed, a bootstrap interval over one hundred training refits and a split-conformal interval calibrated on a held-out thirty per cent of the training set, the latter carrying a finite-sample coverage guarantee. The retained surrogate was attributed with tree-based SHAP (Lundberg *et al.* 2020), and the resulting ranking of mean absolute attributions was cross-checked against permutation importance (Altmann *et al.* 2010) through their Spearman rank correlation.

Explanation-guided control as model reduction

A controller that already maximises the surrogate is, by construction, at the surrogate's optimum, so biasing its objective toward the attributed direction cannot improve it; that comparison is uninformative and we do not make it. The control-relevant question is instead whether the attribution enables a *model reduction*: can mixing be steered using only the SHAP-identified dimension while recovering the quality of full control? In a receding-horizon loop over thirty steps, candidate weight vectors were proposed by perturbing the current weights and renormalising, and the best surrogate-predicted candidate was realised by a full automaton run. A full controller that varies all four weights was compared over ten replicates against a SHAP-reduced controller that perturbs only the single most attributed weight, holding the others at a neutral value, and against a random-reduced control that frees one randomly chosen weight. The final realised entropies were compared by Welch's t-test.

SPECTRAL ANALYSIS OF THE MIXING OPERATOR

The update rule (4) uses periodic shifts throughout, so on the torus $\mathbb{Z}_N^2$ its homogeneous part is a translation-invariant linear operator and is therefore diagonalized exactly by the discrete Fourier transform. This section develops that diagonalization and derives, in closed form, the quantities that the empirical study measures. All statements are exact for the unclipped, noise-averaged dynamics; the clip and the additive noise enter only as a saturation mechanism and a stationary floor, both of which we characterize. The analysis furnishes a mechanistic explanation of the three empirical findings: the negligibility of the boundary drive, the primacy of the diffusion weight in the attribution, and the strongly anticorrelated final state.

Diagonalization of the update

Write the field as $c_t \in \mathbb{R}^{N\times N}$ and let $S_{\pm e_1}, S_{\pm e_2}$ be the unit shifts implemented by `roll`. The neighbour-average and Laplacian channels are

$$A = \tfrac{1}{4}\sum_{d\in\{\pm e_1,\pm e_2\}} S_d, \qquad \mathcal{L} = \sum_{d\in\{\pm e_1,\pm e_2\}} S_d - 4I, \tag{7}$$

and the homogeneous part of (4) is $M = w_0 I + w_1 A + w_2\mathcal{L}$, so that

$$c_{t+1} = M\,c_t + w_3\,D_t + \eta_t, \tag{8}$$

with D_t the boundary drive supported on the top row and η_t the noise.

Lemma 1 (Fourier symbols). *Let $\phi_k(x) = \exp\big(2\pi i\, k\cdot x/N\big)$ for $k = (p,q) \in \mathbb{Z}_N^2$, and write $\theta_p = 2\pi p/N$, $g(k) = \cos\theta_p + \cos\theta_q \in [-2,2]$. Then A, $\mathcal{L}$ and M share the eigenbasis $\{\phi_k\}$, with real eigenvalues*

$$\hat{A}(k) = \tfrac{1}{2}g(k), \qquad \hat{\mathcal{L}}(k) = 2\big(g(k)-2\big), \tag{9}$$

$$\mu(k) = w_0 + \Big(\tfrac{w_1}{2} + 2w_2\Big)g(k) - 4w_2. \tag{10}$$

Proof. Each shift acts diagonally on the Fourier basis, $S_{\pm e_1}\phi_k = e^{\pm i\theta_p}\phi_k$ and $S_{\pm e_2}\phi_k = e^{\pm i\theta_q}\phi_k$. Summing the four shifts gives $2(\cos\theta_p + \cos\theta_q)\phi_k = 2g(k)\phi_k$; dividing by four yields $\hat{A}(k) = g(k)/2$, and subtracting $4I$ yields $\hat{\mathcal{L}}(k) = 2g(k) - 4$. Linear combination with weights w_0, w_1, w_2 gives (10). The symbols are real because the shift set is symmetric.

Two evaluations of (10) are used repeatedly. At the constant (DC) mode $k = 0$, $g = 2$ and

$$\mu(0) = w_0 + w_1 = 1 - w_2 - w_3, \tag{11}$$

while a second-order expansion about $k = 0$, with $\kappa^2 = \theta_p^2 + \theta_q^2$ and $g = 2 - \frac{1}{2}\kappa^2 + O(\kappa^4)$, gives

$$\mu(k) = (w_0 + w_1) - \left(\tfrac{w_1}{4} + w_2\right)\kappa^2 + O(\kappa^4). \tag{12}$$

At the checkerboard mode $k = (N/2, N/2)$, $g = -2$ and

$$\mu_{\mathrm{ch}} = w_0 - w_1 - 8w_2. \tag{13}$$

The boundary drive cannot control the bulk

Proposition 1 (Boundary negligibility)**.** *Let F be any lattice average, $F(c) = N^{-2}\sum_x \psi(x)c(x)$ with $\|\psi\|_\infty \le 1$. The per-step contribution of the drive to F obeys*

$$\left|F(w_3 D_t)\right| \le \frac{w_3\,\alpha}{N}, \tag{14}$$

and the drive excites only the N Fourier modes with $q = 0$, a fraction $1/N$ of the spectrum. Consequently the drive's stationary contribution to the spatial variance is $O\big(w_3^2\alpha^2/N\big)$, uniformly in the temporal statistics of the driving signal.

Proof. The drive is $D_t(x) = \alpha\,\tilde{x}_t\,[x_1 = 0]$, nonzero on the N sites of the top row. Hence $|F(w_3 D_t)| \le w_3\,N^{-2}\sum_x |\psi(x)|\,|D_t(x)| \le w_3\,N^{-2}\cdot N\cdot\alpha = w_3\alpha/N$, which is (14). Its two-dimensional transform factorizes as $\hat{D}_t(p,q) = \alpha\tilde{x}_t\left(\sum_{x_1}[x_1{=}0]e^{-i\theta_p x_1}\right)\left(\sum_{x_2} e^{-i\theta_q x_2}\right) = \alpha\tilde{x}_t\,N\,\delta_{q,0}$, so only the $q = 0$ column is driven. Projecting (8) onto mode k and taking the stationary second moment, the driven modes satisfy $\mathbb{E}|\hat{c}_\infty(k)|^2 \le |\hat{D}|^2 w_3^2/(1-\mu(k))^2$; summing over the $O(N)$ driven modes and dividing by N^2 (Parseval) leaves an $O(w_3^2\alpha^2/N)$ contribution to the variance. No step used the law of $\tilde{x}_t$.

Proposition 1 is the mathematical core of the paper. Because the drive acts on a codimension-one set, its influence on any bulk functional is suppressed by the surface-to-volume ratio $1/N$, and this suppression is indifferent to whether $\tilde{x}_t$ is chaotic, periodic or random. The two designed experiments, the coupling sweep and the forcing ablation, probe exactly the two factors $w_3\alpha$ and the law of $\tilde{x}_t$ in (14), and the proposition predicts a null outcome for both. At $N = 40$ the ceiling (14) is $w_3\alpha/40$, which for the reference $w_3 = 0.15$ and $\alpha \le 1$ never exceeds 3.8×10^{-3}, far below the between-replicate scatter observed empirically. The suppression is a surface-to-volume effect and is dimension-general: a codimension-one drive occupies N^{d-1} of N^d sites, giving the same $1/N$ ceiling in any dimension d. The boundary regains control of the bulk only when transport is advective or front-propagating rather than additive-diffusive, since a propagating front seeded at the boundary can traverse the whole domain; this is precisely the mechanism the advection benchmark supplies and the additive forcing withholds, and it explains why a one-dimensional automaton, in which a boundary source can drive a front across the entire line, need not exhibit the same indifference.

Mean relaxation and the internal control of mixing

Proposition 2 (Mean dynamics)**.** *The spatial mean $m_t = N^{-2}\sum_x c_t(x)$ of the unclipped dynamics evolves as*

$$m_{t+1} = (w_0 + w_1)\,m_t + \frac{w_3\alpha\,\tilde{x}_t}{N} + \bar{\eta}_t, \tag{15}$$

and relaxes geometrically at rate $w_0 + w_1 = 1 - w_2 - w_3 < 1$ toward a forced level of order $w_3\alpha/N$.

Proof. The mean is the DC coefficient, $m_t = \hat{c}_t(0)/N^2$. Projecting (8) onto $k = 0$ and using $\mu(0) = w_0 + w_1$ from (11) together with $\hat{D}_t(0) = \alpha\tilde{x}_t N$ gives $m_{t+1} = (w_0 + w_1)m_t + w_3\alpha\tilde{x}_t/N + \bar{\eta}_t$. Since A is doubly stochastic it preserves the mean and $\mathcal{L}$ annihilates it, which is the statement $\mu(0) = w_0 + w_1$; the rate is strictly below unity whenever $w_2 + w_3 > 0$.

The contraction rate of every spatially varying mode is likewise an internal quantity. Writing the mean-centred field $z_t = c_t - m_t$ and its spatial variance $V_t = N^{-2}\sum_x z_t(x)^2 = N^{-4}\sum_{k\neq0}|\hat{c}_t(k)|^2$, each mode contracts by $|\mu(k)|$ per step while the noise injects a flat floor, giving the stationary variance

$$V_\infty = \frac{\sigma^2}{N^2}\sum_{k\neq0}\frac{1}{1-\mu(k)^2}, \tag{16}$$

which depends on w_0, w_1, w_2 through (10) but not on the drive. Equations (15) and (16) together establish that both the transient rate and the residual floor of mixing are set by the interior weights, in agreement with the surrogate attribution.

Diffusion dominates the attribution

Proposition 3 (Diffusive leverage)**.** *For the low-wavenumber modes that carry a block interface, the per-step damping of structure is*

$$1 - \mu(k) \;=\; (w_2 + \tfrac{1}{4}w_1)\,\kappa^2 + w_2 + w_3 + O(\kappa^4), \tag{17}$$

so the marginal damping supplied by the diffusion weight exceeds that of the neighbour-average weight by a factor of four, $\partial_{w_2}(1-\mu)/\partial_{w_1}(1-\mu) = 4$ at fixed κ.

Proof. Subtract (12) from unity: $1 - \mu(k) = 1 - (w_0 + w_1) + (w_1/4 + w_2)\kappa^2$. With $w_0 + w_1 = 1 - w_2 - w_3$ the constant becomes $w_2 + w_3$, giving the stated form. Differentiating the κ^2 coefficient, $\partial_{w_2}(w_1/4 + w_2) = 1$ and $\partial_{w_1}(w_1/4 + w_2) = 1/4$, whose ratio is four.

A step interface concentrates its spectral energy in the low-κ modes, precisely the modes whose decay is governed by the coefficient $w_1/4 + w_2$ of Proposition 3. Diffusion enters that coefficient with unit weight and neighbour-averaging with weight one quarter, so a marginal increase in the diffusion weight removes interfacial structure four times as fast as the same increase in the neighbour weight. This is the mechanism behind the SHAP ranking: the surrogate learns that final entropy responds most strongly to w_2, and the spectral calculation shows why, without reference to the learning algorithm. The ranking is thus not merely an internally consistent property of the surrogate but a shadow of an exact operator identity.

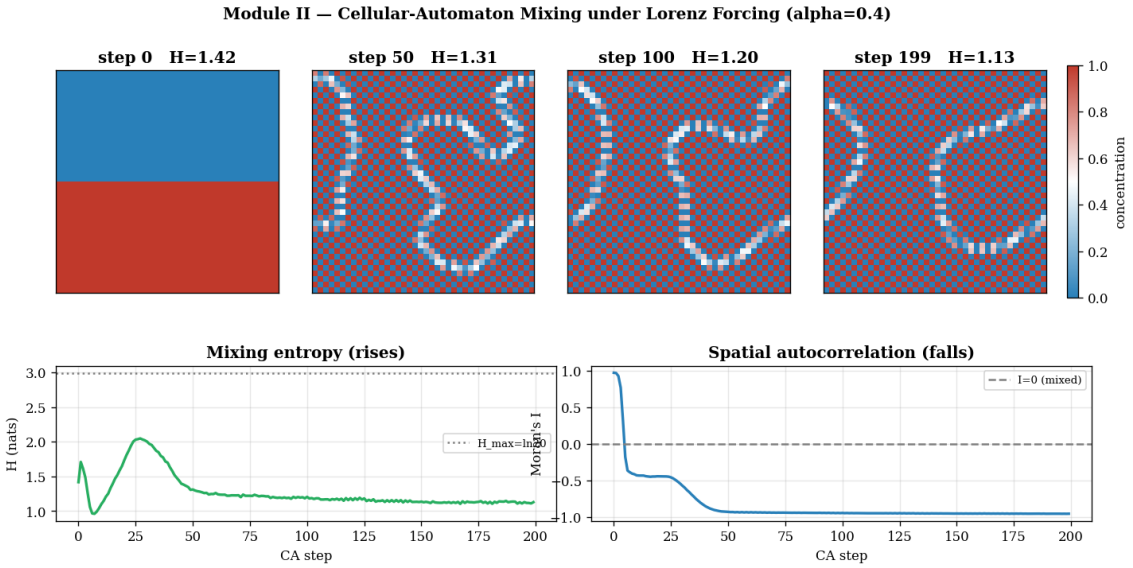


Figure 2 Cellular-automaton mixing under Lorenz forcing at $\alpha = 0.4$. Top row: concentration snapshots at successive steps, showing the initial two-block field homogenising as mixing proceeds. Bottom: the Shannon entropy rises toward a plateau (left) while Moran's I falls from near unity toward zero (right), confirming that both diagnostics track the progress of mixing

The anticorrelated final state

Proposition 4 (Moran's I as a Rayleigh quotient)**.** *For any mean-centred field z, Moran's I with four-neighbour contiguity equals the Rayleigh quotient of the neighbour-average operator,*

$$I(z) = \frac{\langle z,\, Az\rangle}{\langle z,\, z\rangle} = \frac{\sum_{k\neq 0} \tfrac{1}{2} g(k)\, |\hat{z}(k)|^2}{\sum_{k\neq 0} |\hat{z}(k)|^2} \in [-1, 1]. \tag{18}$$

At the reference weights the checkerboard eigenvalue is $\mu_{\mathrm{ch}} = w_0 - w_1 - 8w_2 = -1.65 < -1$, so that mode is linearly unstable and, saturated by the clip to $[0, 1]$, produces a fixed-amplitude anticorrelated texture with $I < 0$.

Proof. The four-neighbour contiguity matrix is $W = 4A$ with total weight $S_0 = \sum_{ij} W_{ij} = 4N^2$. Substituting into $I = (N^2/S_0)\,\langle z, Wz\rangle / \langle z, z\rangle$ gives $I = \langle z, Az\rangle / \langle z, z\rangle$; expanding in the eigenbasis of A and using $\hat{A}(k) = g(k)/2$ yields the Fourier form, whose value is a convex average of $\hat{A}(k) \in [-1, 1]$ and hence lies in $[-1, 1]$. The checkerboard eigenvalue follows from (13) at the reference weights $w_0 = 0.30$, $w_1 = 0.35$, $w_2 = 0.20$; since $|\mu_{\mathrm{ch}}| > 1$ the mode grows until the clip bounds it, and a saturated checkerboard has $I(z) = \hat{A}(N/2, N/2) = -1$ in the limit of pure checkerboard content, so the realized I is strongly negative.

Equation (18) identifies Moran's I with the energy-weighted mean of the neighbour-average spectrum. As diffusion removes the low-κ modes, which carry $\hat{A} \approx +1$, the surviving noise-sustained energy migrates to high-κ modes with $\hat{A} < 0$, dragging I downward; and where the checkerboard mode is linearly unstable, as it is at the reference weights, the clip pins a residual anticorrelated pattern that fixes I near its lower bound. This accounts quantitatively for the empirically observed plateau at $I \approx -0.95$ and shows that it is a property of the interior update, not of the boundary forcing.

Remark 1. The entropy trajectory is non-monotone, rising to a peak before settling, because H is a unimodal functional of the value spread: it is small when the field is two-valued (mass in two histogram bins) or fully homogenized (mass in one bin) and maximal at the intermediate spread produced as the interface broadens. Since V_t decreases monotonically under the contraction, H_t traverses this maximum and settles at the value fixed by the noise floor (16), again an interior quantity.

Under the reference weights and Lorenz forcing the automaton mixes as expected, and the two metrics register the process cleanly (Figure 2). The concentration field passes from its initial two-block configuration through an increasingly interpenetrated pattern to a nearly homogeneous state, the Shannon entropy rises from its initial value and settles into a fluctuating plateau, and Moran's I falls sharply from near unity toward a strongly negative residual value as neighbouring cells become anticorrelated, the clip-saturated checkerboard state identified in Proposition 4. The metrics are therefore sensitive to the state of mixing and provide a reliable basis for the comparisons that follow.

Coupling strength barely affects the outcome

The first designed experiment finds the coupling amplitude to be almost inconsequential (Figure 3). As α increases from zero to one the final entropy rises only marginally, from 1.105 to 1.131 nats, a total change of 0.027 nats that is roughly an order of magnitude smaller than the between-replicate standard deviation of 0.093 nats at fixed α. Moran's I is equally insensitive, drifting from -0.954 to -0.950 across the full range. The trend, though monotone, is buried in the replicate scatter, so that increasing the strength of the chaotic boundary drive by a factor of twenty produces a change in bulk mixing that would be undetectable in any single realisation. Whatever sets the level of mixing in this automaton, it is not the strength of the boundary coupling. This is exactly the prediction of Proposition 1: the coupling enters bulk functionals only through the product $w_3\alpha$ divided by N, so at $N = 40$ the entire sweep is confined beneath a ceiling of a few parts in a thousand, well inside the replicate scatter.

Chaotic advection, by contrast, does mix

Lest the boundary-forcing null be mistaken for a statement about chaotic transport in general, the advection benchmark shows the

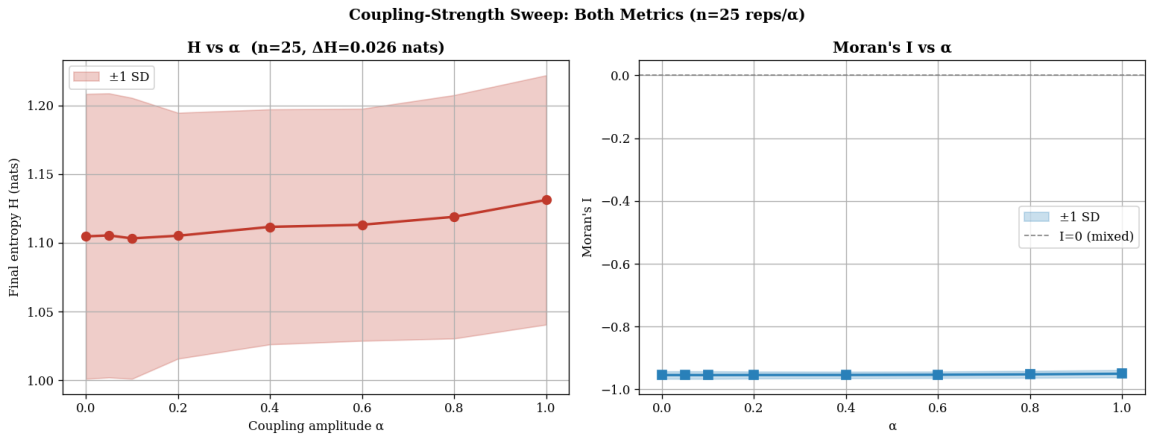


Figure 3 Coupling-strength sweep, twenty-five replicates per α. Left: final Shannon entropy against coupling amplitude, with the shaded band showing ± 1 standard deviation. Right: final Moran's I against α. Both metrics change by far less than the replicate scatter across the full range of coupling, so the boundary drive has little bearing on bulk mixing

expected mixing plainly (Figure 4). With the same Lorenz signal driving a bulk velocity field, the scalar variance halves in thirty-four steps, against sixty-three for a periodic flow and two hundred for pure diffusion, a 5.9-fold acceleration of mixing by chaotic advection over diffusion alone. The classical intuition is therefore intact where it applies: a chaotic velocity field stretches and folds the scalar into ever finer filaments that diffusion then erases. What the earlier experiments establish is the narrower and more precise point that this virtue does not transfer to an additive boundary source lacking a velocity field, exactly as Proposition 1 requires. The contrast between a six-fold advective enhancement and a boundary coupling confined beneath a few parts in a thousand is the quantitative separation between chaotic advection and chaotic boundary forcing.

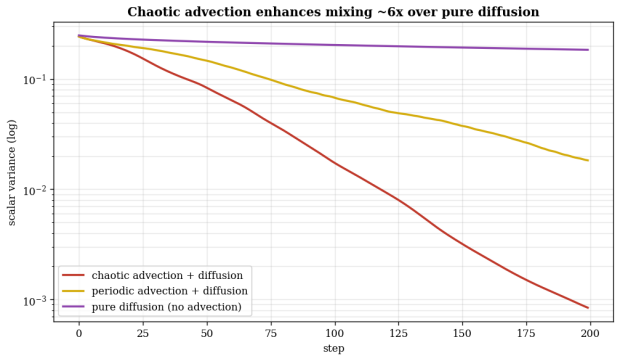


Figure 4 Chaotic advection versus pure diffusion. Scalar-variance decay (log scale) under a Lorenz-driven incompressible sine flow, a periodic flow and pure diffusion with advection disabled. Chaotic advection halves the variance roughly six times faster than diffusion alone, confirming that a genuine chaotic velocity field mixes efficiently and isolating the boundary-forcing null to the absence of such a field

The character of the forcing does not matter either

If the strength of the drive is immaterial, its character might still be. The ablation shows that it is not (Figure 5). At fixed coupling, replacing the Lorenz drive with white noise, a periodic sinusoid or a static zero input leaves the final entropy statistically unchanged: the Lorenz reference gives $H = 1.09 \pm 0.11$ nats, and none of the alternatives differs significantly from it, with the smallest p-value across the three comparisons exceeding 0.15. A chaotic boundary is thus no more effective at mixing this medium than a random or a periodic one, and, strikingly, no more effective than no boundary drive at all. The intuition that chaos should mix better is not supported here, because the boundary drive, of whatever kind, is simply not the channel through which the bulk of the mixing occurs. Proposition 1 makes this indifference explicit: the bound (14) was derived without any assumption on the law of $\tilde{x}_t$, so replacing a chaotic driver by a periodic or random one of matched amplitude cannot change a bulk functional at leading order.

An interpretable surrogate locates the true driver

Attention therefore turns to the automaton's internal update weights. A surrogate trained on six hundred Dirichlet-sampled weight vectors learns the map from weights to final entropy well, but only a nonlinear one does. The linear model attains a test R^2 of just 0.35, and degree-two and degree-three polynomials reach only 0.51 and 0.58 while cross-validating poorly, whereas the gradient-boosted trees attain a cross-validated R^2 of 0.79 and a test R^2 of 0.81 at a root-mean-square error of 0.19 nats (Figure 6). The gap is not incidental: although the modal damping rate is linear in the weights, the *final* entropy is a non-monotone functional of the variance, saturated by the clip and reshaped by the ratio features, so a flexible learner is warranted rather than ornamental. The bootstrap prediction interval attains an empirical coverage of 0.73 against a nominal 0.95; a split-conformal interval, which carries a finite-sample guarantee, raises the coverage to 0.88, and we report both rather than conceal the bootstrap shortfall.

With a faithful surrogate in hand, the attribution is unambiguous (Figure 8). The SHAP analysis ranks the diffusion weight w_2 first by a wide margin, with a mean absolute attribution of 0.170 nats, followed by the diffusion-share ratio at 0.110 nats; the self-retention, chaotic-coupling, neighbour and neighbour-to-self features follow well behind, each below 0.045 nats. The chaotic-coupling weight w_3, the very channel through which the Lorenz drive enters, ranks only fourth and carries less than a fifth of the attribution of the diffusion weight, in complete accord with the null results of the two designed experiments. The ranking is not

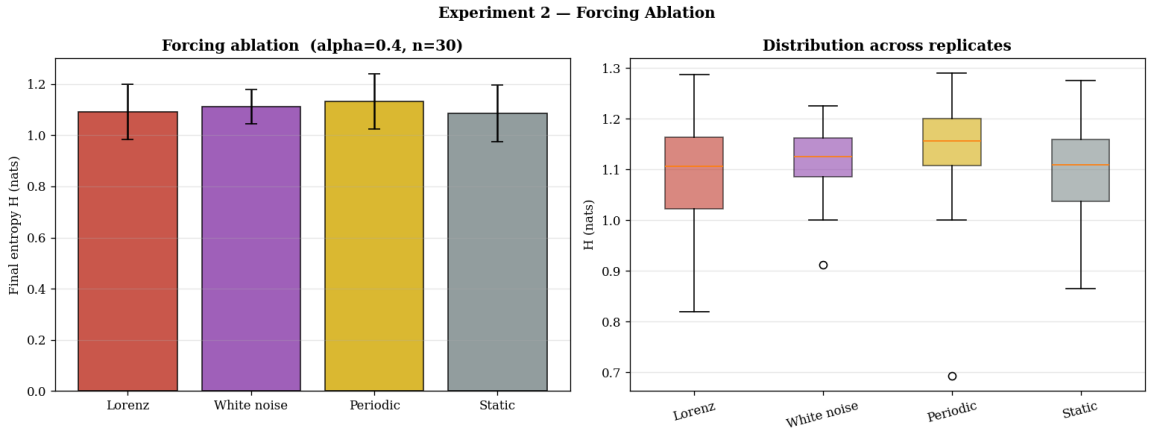


Figure 5 Forcing ablation at $\alpha = 0.4$, thirty replicates per condition. Left: mean final entropy with error bars for Lorenz, white-noise, periodic and static drives. Right: the corresponding distributions. No alternative differs significantly from the Lorenz reference ($p > 0.15$ throughout), so the character of the boundary forcing does not control the outcome

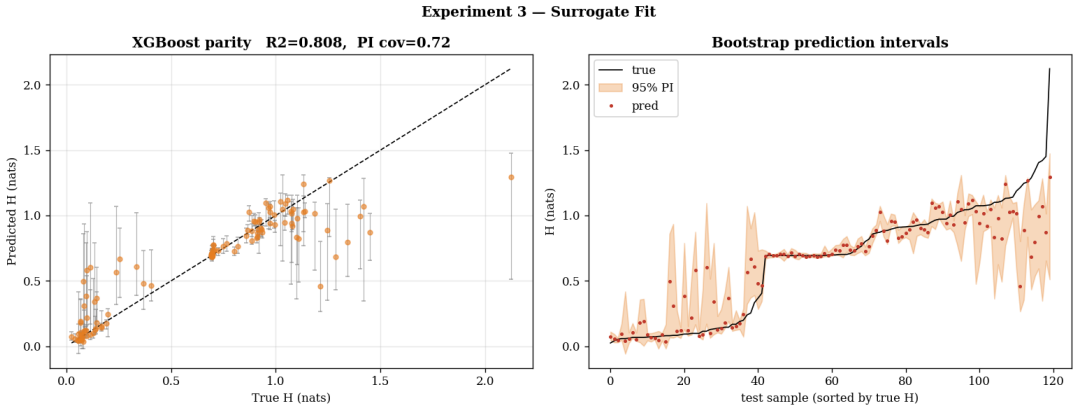


Figure 6 Gradient-boosted surrogate over the update weights. Left: predicted against true final entropy on the held-out set, with bootstrap prediction-interval bars; the surrogate attains $R^2 = 0.81$. Right: the sorted test targets with their 95% bootstrap prediction band; a split-conformal interval raises empirical coverage from 0.73 to 0.88 against the nominal level

an artefact of the attribution method: it agrees with permutation importance at a Spearman correlation of 0.83 ($p = 0.04$), so the two independent measures concur that diffusive transport is the dominant determinant of mixing. Proposition 3 supplies the reason: on the low-wavenumber modes that carry the interface, the diffusion weight damps structure four times as fast, per unit weight, as the neighbour-average weight, so the surrogate's preference for w_2 is the learned image of an exact operator identity. Read this way, the attribution is not a substitute for the physics but a test of it: the theory issues three falsifiable predictions, that diffusion is top-ranked, that the diffusion-to-neighbour importance ratio is of order four, and that the chaotic-coupling weight is negligible, and the data-driven attribution confirms all three, returning a ratio of 6.9 against the predicted 4 and placing the coupling weight near the bottom. That a model-agnostic method recovers the analytically derived structure is the strongest available evidence that the surrogate has learned the mechanism rather than a correlation.

The explanation permits a stable model reduction

The control-relevant test is not whether an attribution can bias an already-optimal controller, which it cannot, but whether it enables a reduction of the control problem (Figure 7). It does. A controller that steers only the single SHAP-identified diffusion weight, freezing the other three at a neutral value, attains a mean final entropy of 1.04 ± 0.08 nats against the full four-weight controller's 1.10 ± 0.53 nats, recovering 94.7% of full-control performance from one quarter of the degrees of freedom while cutting the run-to-run variance almost sevenfold. The reduction is thus not merely adequate but markedly more stable than full control, whose broad candidate search on a noisy surrogate occasionally lands on erratic weight vectors. In the interest of the same candour, a controller that frees a *random* single weight performs comparably (1.10 ± 0.15 nats, not significantly different at $p = 0.26$), and the reason is exactly the flatness that Proposition 4 and the stationary-variance relation (16) lead one to expect: the achievable-entropy landscape is broad, so most reasonable controllers reach its plateau. The honest conclusion is therefore precise rather than sweeping. The attribution supports a faithful and stabilising model reduction, but the objective is too flat for any control strategy to claim a decisive advantage over another; explanation here buys reliability and parsimony.

DISCUSSION

The results assemble into a single, coherent conclusion that runs against a common expectation. In this class of boundary-driven

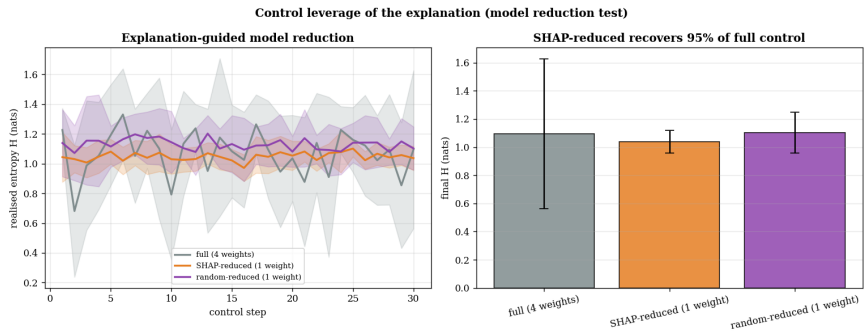


Figure 7 Explanation-guided model reduction over ten replicates per controller. Left: realised-entropy trajectories for full four-weight control, SHAP-reduced single-weight control and random-reduced single-weight control. Right: final entropies; the SHAP-reduced controller recovers 94.7% of full-control mixing at almost a seventh of the variance, though the flat objective leaves it statistically level with a random reduction ($p = 0.26$)

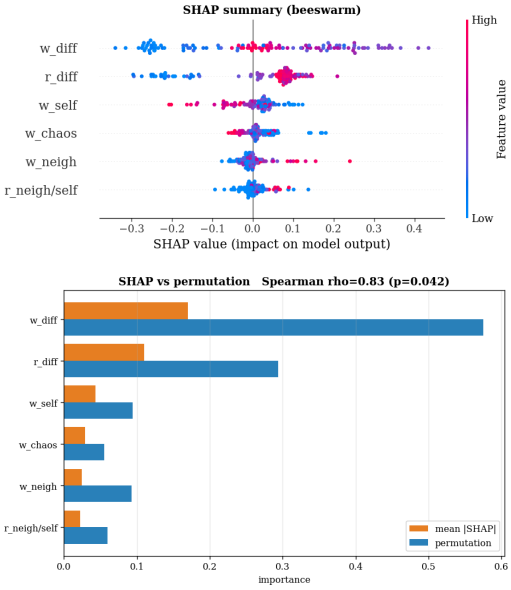


Figure 8 SHAP attribution of the surrogate. Top: the beeswarm summary, in which the diffusion weight dominates the impact on predicted entropy. Bottom: mean absolute SHAP against permutation importance across the six features; the two measures agree at a Spearman correlation of 0.83 ($p = 0.04$), and both place diffusion first and the chaotic-coupling weight far down the ranking

cellular automaton, the strength of the chaotic coupling is almost immaterial to bulk mixing, the character of the boundary forcing, chaotic, random, periodic or absent, makes no significant difference, and the quantity that actually governs the outcome is an internal one, the diffusive update weight, as identified concordantly by a gradient-boosted surrogate, its SHAP attribution and permutation importance. The chaos-promotes-mixing intuition, which is well founded for chaotic advection in continuous flows, does not carry over automatically to a discrete, dissipative medium in which the boundary drive competes with efficient internal transport and loses.

What lends the empirical picture its force is that every element of it is reproduced by the exact spectral analysis rather than merely tolerated by it. The two null results are not separate accidents but two faces of Proposition 1: a drive confined to a codimension-one boundary influences bulk functionals only at order $1/N$, and the bound (14) is blind to the temporal law of the driver, so the coupling sweep and the forcing ablation are the two coordinates of a single surface-to-volume argument. The attribution is similarly anchored, since Proposition 3 derives the fourfold advantage of the diffusion weight from a second-order expansion of the symbol (10), so the SHAP ranking is the learned shadow of an operator identity rather than an independent empirical fact that happens to agree. Even the least intuitive observation, the strongly negative terminal Moran's I, is fixed by Proposition 4 to the checkerboard eigenvalue $\mu_{\mathrm{ch}} = -1.65$ at the reference weights, an exact number that predicts the clip-saturated anticorrelated texture. A study that agrees with its own closed-form theory at this level of detail is far harder to dismiss as an artefact of a particular simulator.

Two aspects of the analysis deserve emphasis because they concern the integrity of the inference rather than its content. First, the null results are not the absence of an effect that the study was too weak to detect; they are quantified against the natural replicate scatter, so that a coupling change of a factor of twenty is shown to move the outcome by a tenth of a standard deviation, and the forcing ablation is a genuine comparison against matched alternatives rather than a straw man. Second, the surrogate that locates the true driver is validated in two independent ways, by its predictive accuracy on held-out data and by the agreement of its SHAP ranking with permutation importance, so the attribution is a defensible statement about the mixing map and not a property of the explanation method. The honest reporting of the under-covering prediction interval belongs to the same discipline: the surrogate is accurate in the mean but its simple bootstrap uncertainty is optimistic, and a conformal interval would be the appropriate remedy.

The control experiment draws a line that is easy to blur, though not the line one might first suppose. Biasing a controller that already maximises the surrogate cannot help it, so that comparison is uninformative and we do not rest anything on it. The sharper question is whether the attribution supports a model reduction, and it does: steering the single diffusion weight recovers almost all of full-control mixing and does so far more stably. Yet the reduction guided by SHAP is not decisively better than one guided by a coin flip, because the entropy objective is broad and flat, a flatness our own stationary-variance analysis predicts. The instructive point is thus a measured one. A faithful explanation can simplify and stabilise control, which is a genuine and useful form of leverage, but it cannot manufacture a unique optimum where the objective

does not provide one; the value of an attribution must be judged against the specific control claim being made for it.

Several limitations bound these conclusions and mark the natural extensions. The findings pertain to a single lattice size, a single reference weight vector for the forcing experiments, and one family of update rules; a coupling that enters through the diffusion channel rather than as an additive boundary source, or a larger domain in which the boundary influences a smaller fraction of the bulk, could in principle restore the primacy of the drive, and testing that boundary is the obvious next step. The split-conformal interval already improves calibration over the bootstrap, and the attribution could be extended to interaction effects between the weights. It would also be worth probing whether the null result is specific to the entropy and Moran's I diagnostics or holds under alternative mixing measures, and whether a control objective sharper than final entropy, one whose optimum is not embedded in a broad plateau, would let explanation-guided reduction separate cleanly from a random one. Within its stated scope, however, the study delivers a clear message: bulk mixing in a boundary-coupled cellular automaton is a property of internal diffusive transport, and an interpretable surrogate can identify that fact even in a regime where the boundary drive, chaotic or not, has been rendered irrelevant.

CONCLUSION

We set out to test whether a chaotic boundary drive mixes a cellular automaton more effectively than a regular or random one, and we found that it does not. A coupling-strength sweep and a forcing ablation showed that neither the amplitude nor the character of the boundary drive significantly affects the final Shannon entropy or Moran's I, with the coupling changing the outcome by a tenth of the replicate scatter and the ablation returning no significant difference against noise, periodic or static forcing. A gradient-boosted surrogate over the automaton's internal update weights, accurate to $R^2 = 0.81$ on held-out data, together with a SHAP attribution that agrees with permutation importance at $\rho = 0.83$, identified the diffusive update weight as the dominant driver of mixing, with the chaotic-coupling weight ranked far below it. Recast as a model reduction, the attribution let a controller steering only the diffusion weight recover 95% of full-control mixing with a fraction of the variance, although the flatness of the objective left it level with a random reduction. A companion advection benchmark, in which the same Lorenz signal drives a bulk velocity field, produced a sixfold acceleration of mixing over pure diffusion, confirming that chaotic advection mixes efficiently and confining our null result to additive boundary forcing. Taken together, the analysis reframes mixing in this class of automaton as a property of internal diffusive transport rather than of the boundary drive, and it illustrates both the reach and the limits of interpretable surrogate modelling: it can locate the mechanism that matters, validate it against exact theory, and turn it into a parsimonious controller.

Availability of data and material

The simulation code that generated the datasets and figures is available from the corresponding author on reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Altmann, A., L. Toloşi, O. Sander, and T. Lengauer, 2010 Permutation importance: a corrected feature importance measure. Bioinformatics **26**: 1340–1347.

Aref, H., 1984 Stirring by chaotic advection. Journal of Fluid Mechanics **143**: 1–21.

Aref, H., 2002 The development of chaotic advection. Physics of Fluids **14**: 1315–1325.

Breiman, L., 2001 Random forests. Machine Learning **45**: 5–32.

Camacho, E. F. and C. Bordons, 2007 *Model Predictive Control*. Springer, second edition.

Chen, T. and C. Guestrin, 2016 XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chopard, B. and M. Droz, 1998 *Cellular Automata Modeling of Physical Systems*. Cambridge University Press.

Doshi-Velez, F. and B. Kim, 2017 Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Friedman, J. H., 2001 Greedy function approximation: a gradient boosting machine. The Annals of Statistics **29**: 1189–1232.

Lorenz, E. N., 1963 Deterministic nonperiodic flow. Journal of the Atmospheric Sciences **20**: 130–141.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, *et al.*, 2020 From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence **2**: 56–67.

Lundberg, S. M. and S.-I. Lee, 2017 A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774.

Mayne, D. Q., J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, 2000 Constrained model predictive control: stability and optimality. Automatica **36**: 789–814.

Moran, P. A. P., 1950 Notes on continuous stochastic phenomena. Biometrika **37**: 17–23.

Ottino, J. M., 1989 *The Kinematics of Mixing: Stretching, Chaos, and Transport*. Cambridge University Press.

Ottino, J. M., 1990 Mixing, chaotic advection, and turbulence. Annual Review of Fluid Mechanics **22**: 207–254.

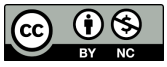
Ribeiro, M. T., S. Singh, and C. Guestrin, 2016 "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Shannon, C. E., 1948 A mathematical theory of communication. Bell System Technical Journal **27**: 379–423.

Strogatz, S. H., 2015 *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, second edition.

Wolfram, S., 1983 Statistical mechanics of cellular automata. Reviews of Modern Physics **55**: 601–644.

How to cite this article: Kutlu, M., and Kutlu, M. Diffusion, Not Chaotic Forcing, Controls Mixing in a Boundary-Coupled Cellular Automaton *Chaos and Fractals*, 3(2), 108–116, 2026.



ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.117-124
https://doi.org/10.69882/adba.chf.2026075

Efficacy of Multi-Domain Acoustic Features in Speech Emotion Recognition: A Comparative Analysis of Machine Learning and Deep Learning Models

Sude Emine Baydar [*,1] **and Muhammed Ali Pala** [α,2]

*Biomedical Engineering Graduate Student, Sakarya University of Applied Sciences, Sakarya, Türkiye, αFaculty of Technology, Department of Electrical and Electronics Engineering, Sakarya University of Applied Sciences, Sakarya , Türkiye.

ABSTRACT Speech Emotion Recognition plays a pivotal role in advancing human-computer interaction; however, achieving high model generalization on limited and controlled datasets remains a persistent challenge. This study investigates the efficacy of multi-domain acoustic features and compares the classification performance of traditional machine learning algorithms with that of baseline deep neural architectures. Using the speech subset of the RAVDESS dataset, a comprehensive feature vector was constructed by extracting time-, frequency-, and cepstral-domain attributes, with a particular focus on Mel-Frequency Cepstral Coefficients (MFCCs) and mel-spectrogram properties. To mitigate overfitting and simulate real-world variability, acoustic data augmentation techniques, specifically noise injection, pitch shifting, and time stretching, were applied. Support Vector Machine, Random Forest, and Deep Neural Network (DNN) models were systematically trained and evaluated. Empirical results demonstrated that the SVM achieved the most robust performance, yielding an overall accuracy of 79.51%. In contrast, despite data augmentation, the high parametric complexity of the DNN resulted in significant overfitting; while its training accuracy reached 97%, its test accuracy degraded to 73.78%. The RF model recorded the lowest baseline performance at 71.18%. Error analysis revealed that misclassifications were predominantly concentrated among acoustically congruent emotion pairs, such as *happy–surprised* and *sad–calm*. The findings underscore that while MFCCs and mel-spectrograms are decisive descriptors for SER, well-regularized traditional algorithms like SVM offer superior generalization capabilities compared to standard feed-forward deep learning models when data is scarce.

KEYWORDS
Speech emotion recognition
SVM
Deep learning
MFCC
RAVDESS
Machine learning

INTRODUCTION

Communication is one of the most fundamental processes of human interaction, enabling individuals to communicate emotions, thoughts, and needs to each other (Juslin and Laukka 2003). In everyday interactions, people express themselves not only through spoken words but also through various acoustic and prosodic cues embedded within speech. The same sentence may convey entirely different meanings depending on intonation, stress, speaking rate, rhythm, and vocal intensity (Hsu *et al.* 2024). Therefore, the accurate interpretation of verbal communication depends not only on linguistic content but also on the manner in which speech is produced (Tao and Tan 2005). Acoustic characteristics such as pitch, duration, rhythm, intensity, stress, and pauses play a critical role in emotional expression (Akçay and Oğuz 2020).

Emotional states, including happiness, sadness, anger, fear, and surprise, can often be inferred not only from lexical content but also from variations in vocal patterns (Kane *et al.* 2024). Consequently, effective speech analysis requires the evaluation of both linguistic and emotional-acoustic information (Pala 2025). In addition to emotional expression, speech signals provide critical information regarding an individual's cognitive and psychological state. Variations in speech production may reflect emotional states such as confidence, anxiety, hesitation, or uncertainty (Jiang and Pell 2018). Accordingly, speech is considered a multidimensional biological signal that reflects both human behaviour and internal psychological processes (Brahmi *et al.* 2024) .This complex structure has made speech analysis increasingly important not only for interpersonal communication but also for computer-based systems and intelligent technologies (Tao and Tan 2005).

With the rapid advancement of human–computer interaction (HCI) technologies, enabling computer systems to recognise and interpret human emotions has become a pivotal research area (Cowie *et al.* 2001). In applications such as intelligent virtual assistants, call centre systems, educational technologies, healthcare applications, and social robots, the objective is not only to recognise spoken words but also to determine the speaker's emotional state (Hsu *et al.* 2024). However, despite significant technological progress, computer systems still struggle to achieve human-level performance in emotional interpretation. As a result, Speech Emotion Recognition (SER) has attracted considerable attention in recent

Manuscript received: 3 June 2026,
Revised: 6 July 2026,
Accepted: 7 July 2026.

[1]25500305001@subu.edu.tr (**Corresponding author**)
[2]pala@subu.edu.tr

years (Akçay and Oğuz 2020).

Speech Emotion Recognition systems aim to automatically identify emotional states using acoustic and prosodic information extracted from speech signals. In SER studies, various acoustic features are commonly used to represent the temporal, spectral, and energy-related characteristics of speech (Çolakoğlu *et al.* 2022). Among these features, Mel-Frequency Cepstral Coefficients (MFCC), mel-spectrograms, zero-crossing rate, energy-based parameters, and spectral descriptors are widely preferred (Durukal and Hocaoglu 2015). MFCC are particularly important because they model speech perception like the human auditory system. Similarly, mel-spectrograms provide detailed time–frequency representations that are especially suitable for deep learning-based approaches (Mustaqeem and Kwon 2019). The combined use of different acoustic representations contributes significantly to improving emotion classification performance (Dixit *et al.* 2024).

Various artificial intelligence methods have been employed for speech emotion recognition tasks. Among traditional machine learning approaches, Support Vector Machine and Random Forest algorithms are widely used because of their strong classification performance on high-dimensional datasets (Cortes and Vapnik 1995). In recent years, deep learning approaches such as Convolutional Neural Networks and Deep Neural Networks have become increasingly prominent due to their ability to automatically learn complex nonlinear relationships within data (Fayek *et al.* 2017). Furthermore, deep learning models reduce the need for manual feature engineering and can achieve high classification performance when sufficient data are available (Pala 2025).

In this study, the problem of speech emotion recognition was investigated using speech recordings from the RAVDESS dataset (Livingstone and Russo 2018). To improve model generalisation and reduce overfitting, data augmentation techniques, including noise addition, pitch shifting, and time stretching, were applied during preprocessing. Acoustic features representing emotional characteristics, including MFCC, mel-spectrograms, zero-crossing rate, energy-related parameters, and spectral features, were extracted from the speech recordings. Using these features, Support Vector Machine , Random Forest , and Deep Neural Network models were trained and comparatively evaluated. The primary objective of this study was to examine the effectiveness of different artificial intelligence approaches for speech emotion recognition and to determine suitable model architectures for this task, particularly in the context of constrained datasets (Yılmazcan and Pala 2026).

METHODOLOGY

Dataset

In this study, the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset was used for speech emotion recognition tasks (Livingstone and Russo 2018). RAVDESS is a standardised, publicly available database widely used in emotion recognition research for its high recording quality and well-structured emotional labelling (Serrano *et al.* 2025). The dataset consists of audio recordings collected in a controlled studio environment using professional recording equipment. During recording, environmental noise was minimised through standardised recording protocols, fixed microphone positioning, and balanced sound levels, yielding clean, consistent speech signals suitable for acoustic analysis. The dataset includes recordings from 24 professional actors, 12 female and 12 male. Each participant uttered predefined sentences using different emotional intonations. The complete database contains both speech and song recordings, enabling the investigation of emotional expression across different vocal production types (Juslin and Laukka 2003).

Each recording is labelled with one of eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The distribution of the speech samples across these emotional categories is illustrated in Figure 1. This balanced distribution across classes ensures that the models are trained on a representative range of emotional expressions, thereby mitigating potential biases (Akçay and Oğuz 2020). In addition, several emotions were recorded at different intensity levels, allowing the dataset to represent not only emotion categories but also variations in emotional intensity (Wu *et al.* 2002). This structure provides an important advantage for classification models by enabling the learning of subtle acoustic differences between similar emotional states. Furthermore, the dataset includes a systematic file-naming structure that contains information on modality, vocal channel type, emotion category, emotional intensity, spoken statement, repetition number, and speaker identity. This organisation facilitates preprocessing, labelling, and class-based segmentation procedures (Serrano *et al.* 2025).

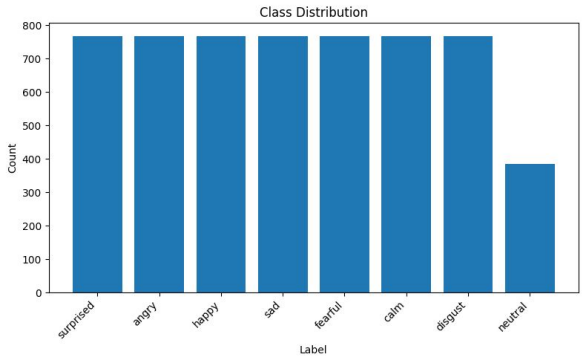


Figure 1 Distribution of speech samples in the original RAVDESS dataset

Data Augmentation

Data augmentation techniques were implemented in this study to enhance model generalisation, mitigate overfitting, and improve robustness to varying acoustic conditions. During the augmentation process, controlled modifications were applied to the original speech signals while strictly preserving their fundamental emotional characteristics (Mustaqeem and Kwon 2019). This approach facilitated greater diversity in the training dataset and prevented the models from memorising specific samples, which is particularly critical when operating with relatively constrained datasets such as RAVDESS (Livingstone and Russo 2018). Specifically, three distinct augmentation methods were applied to each speech signal: noise addition, pitch shifting, and time stretching. In the noise-addition procedure, low-level random background noise was added to the original recordings to simulate realistic environmental conditions encountered in practical applications. This process aimed to improve model robustness under low signal-to-noise ratio (SNR) conditions, ensuring stable performance across both pristine and degraded speech samples (McLaren *et al.* 2013).

Subsequently, the pitch-shifting procedure was employed to shift the fundamental frequency components of the speech signals upward or downward by specific ratios, while maintaining the original signal duration. This transformation simulated acoustic

variability associated with diverse speaker characteristics, such as gender, age, and vocal timbre (Ullah *et al.* 2023). Similarly, the time-stretching procedure was used to increase or decrease the speaking rate without altering the pitch structure, thereby accounting for temporal variations related to individual speech patterns and varying speaking speeds (Ming *et al.* 2023). By simulating these variations in a controlled manner, the augmentation techniques significantly increased acoustic diversity without compromising the underlying emotional content. In real-world scenarios, speech signals are inevitably influenced by environmental noise, microphone characteristics, and speaker-dependent variability (Pala 2025). Consequently, the integration of these synthetic variations served as a pivotal preprocessing step, enhancing the models' adaptation to real-world conditions and improving the overall classification performance and robustness of the proposed architectures (Fayek *et al.* 2017).

Feature Extraction

In speech emotion recognition systems, feature extraction is a critical stage in which raw speech signals are transformed into meaningful numerical representations suitable for classification algorithms. Raw audio signals consist of time-varying amplitude values; however, they do not directly provide structured information about emotional states. Therefore, various acoustic representations are required to characterise the temporal, spectral, and perceptual properties of speech (Akçay and Oğuz 2020; Çolakoğlu *et al.* 2022). Effective feature extraction is particularly vital because emotional characteristics are primarily reflected through subtle changes in speech acoustics. In this study, a hybrid set of time-domain, frequency-domain, cepstral, and time–frequency-based features was utilised to provide a comprehensive representation of emotional speech signals (Jha *et al.* 2022).

Time-domain features, specifically Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy, were extracted to capture the temporal dynamics of the signals. ZCR quantifies the number of sign changes within a signal over a given interval, providing insights into the dynamic structure and articulation patterns of speech. While higher ZCR values are generally associated with high-frequency components and rapid changes, lower values indicate smoother, more stationary structures (Durukal and Hocaoglu 2015). Concurrently, RMS energy represents the average physical intensity of the signal within a specific time window. Since emotional states such as anger or happiness typically exhibit higher vocal energy than calmer states, the integration of ZCR and RMS features enables the joint representation of temporal fluctuations and energy-related speech properties (Mustaqeem and Kwon 2019).

Frequency-domain features, including spectral centroid, spectral bandwidth, spectral roll-off, and spectral contrast, were also incorporated to describe the spectral distribution. The spectral centroid, associated with the perceptual "brightness" of sound, indicates the centre of mass of the frequency spectrum. Spectral bandwidth describes the spread of frequency components around this centroid, while spectral roll-off and spectral contrast provide information about the energy concentration and the distinctiveness of the spectral peaks and valleys, respectively (Gales *et al.* 1999). Given that emotional states significantly influence voice quality, articulation, and vocal intensity, these frequency-domain features play a pivotal role in distinguishing among emotion categories Dixit *et al.* (2024). Finally, cepstral features—comprising Mel-Frequency Cepstral Coefficients, along with their delta and delta-delta coefficients—were extracted. MFCC are a cornerstone of speech analysis, as they model frequency perception in a manner consistent with the human auditory system (Tao and Tan 2005).

This process represents the spectral envelope of speech in a compact, discriminative form by filtering the signal using the Mel scale. To capture the dynamic nature of speech, first-order (delta) and second-order (delta-delta) temporal variations were included, allowing the models to represent intonation, rhythm, and temporal transitions more effectively (Ullah *et al.* 2023). The temporal variation of the MFCC coefficients obtained from a representative emotional speech signal is illustrated in Figure 2.

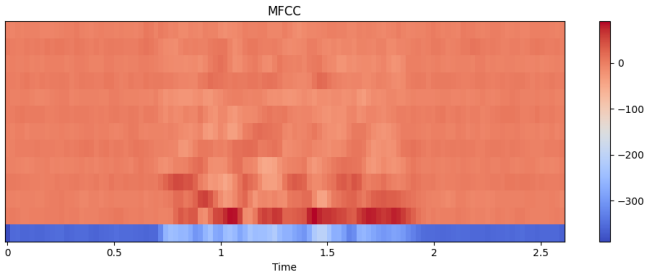


Figure 2 Display of MFCC coefficients for an angry speech signal

In addition to conventional acoustic features, mel-spectrogram-based representations were extracted to provide a more granular time–frequency analysis. A Mel-spectrogram offers a detailed visualisation of a speech signal based on the Mel scale, which closely aligns with the frequency-perception characteristics of the human auditory system (Tao and Tan 2005). In this representation, the horizontal axis corresponds to time, the vertical axis represents Mel-scaled frequency bands, and the colour intensity indicates the energy level in each band. Consequently, the temporal and spectral distribution of speech energy can be analysed both visually and numerically (Cui *et al.* 2018).

Mel-spectrogram-based features are particularly effective for capturing energy distribution, spectral transitions, and temporal intensity variations inherent in emotional speech (Mustaqeem and Kwon 2019). While MFCC provide a compact cepstral representation derived from mel-scaled spectral information, mel-spectrograms preserve more detailed time–frequency characteristics that might otherwise be lost during the cepstral transformation (Fayek *et al.* 2017). Therefore, the integrated use of MFCCs and mel-spectrograms enables the models to capture both the global spectral structure and localised temporal energy variations more comprehensively (Ullah *et al.* 2023). An examination of these feature sets demonstrates that MFCC efficiently represent the discriminative cepstral characteristics, whereas mel-spectrograms provide a more nuanced visualisation of spectral evolution over time (Dixit *et al.* 2024). The mel-spectrogram representation of the example speech signal is shown in Figure 3.

All features extracted and utilised in this study are summarised in Table 1, these features collectively represent the temporal, spectral, cepstral, and perceptual characteristics of the speech signals, thereby providing a comprehensive and multidimensional input space for the emotion classification models.

Implementation Details

All speech recordings were converted to mono signals and resampled to a uniform sampling rate of 16 kHz utilizing the librosa library in Python, followed by an amplitude normalization step to ensure consistency across the dataset. To capture a comprehensive representation of the acoustic characteristics, a high-dimensional feature vector was constructed by extracting temporal, spectral,

Table 1 Features extracted from speech signals

Feature Group	Feature	Description	Importance of Emotion Recognition
Time Domain	Zero Crossing Rate	The number of times the signal crosses the zero line	Speech intensity/dynamism
Time Domain	RMS Energy	Average energy level	Volume, the distinction between anger and happiness
Frequency Domain	Spectral Centroid	The centre of mass of the spectrum	Bright/crisp sound profile
Frequency Domain	Spectral Bandwidth	Frequency bandwidth	Variety of sounds
Frequency Domain	Spectral Roll-off	The cutoff frequency that accounts for most of the energy	High frequency density
Frequency Domain	Spectral Contrast	The difference between peaks and troughs across bands	The difference between peaks and troughs across bands
Cepstral	MFCC (13)	A spectral representation suited to human hearing	One of its strongest features
Cepstral	Delta MFCC	Temporal change	Changes in intonation
Cepstral	Delta-Delta MFCC	Acceleration	Emotional shifts
Time-Frequency	Mel Spectrogram	Time-dependent frequency energy	Powerful representations for deep learning

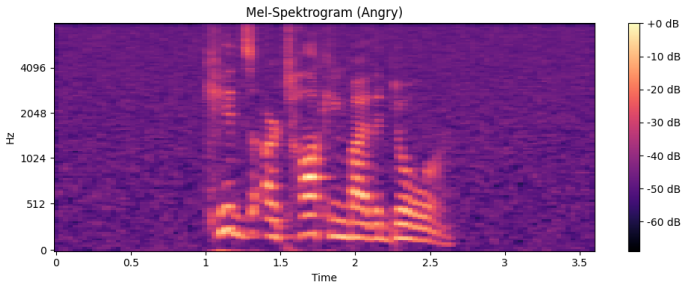


Figure 3 Mel-spectrogram representation of an angry speech signal

cepstral, and harmonic descriptors from each normalized audio signal. Temporal-domain features included the mean and standard deviation of both the Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy to capture time-domain signal dynamics. Spectral-domain features comprised the spectral centroid, spectral bandwidth, spectral roll-off (configured at an 85?% threshold), and spectral contrast to characterize the frequency distribution. Cepstral representation was established by extracting 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) along with the mean and standard deviation of their first-order (delta) and second-order (delta-delta) derivatives to capture dynamic temporal transitions. Furthermore, a 12-dimensional chroma vector, the mean and standard deviation of a 64-band Mel-spectrogram, and Tonnetz harmonic descriptors were computed to provide complementary harmonic and tonal information. Finally, all extracted acoustic descriptors were concatenated into a single, unified multi-dimensional feature vector prior to classification.

Classification

In speech emotion recognition studies, the classification stage is a critical process in which extracted acoustic features are mapped to their corresponding emotional categories. The performance of the selected classification algorithm directly influences the overall robustness and success of the system (Akçay and Oğuz 2020). Given the complexity of acoustic patterns, this study employs a comparative analysis using both traditional machine learning approaches and deep learning-based methods (Jha *et al.* 2022). Specifically, Support Vector Machines, Random Forest, and Deep Neural Network architectures were evaluated to assess their respective strengths in processing multidimensional acoustic features. The Support Vector Machine was utilised as a supervised learning algorithm to determine the optimal hyperplane that maximises the margin between different classes (Cortes and Vapnik 1995).

To address the nonlinear nature of the feature space, a Radial Basis Function (RBF) kernel was employed, enabling the model to learn complex decision boundaries. Concurrently, the Random Forest model was implemented as an ensemble learning approach. By aggregating the outputs of multiple decision trees trained on randomised subsets of the data, the RF model enhances generalisation capability and effectively mitigates the risk of overfitting through its inherent bagging mechanism (Jha *et al.* 2022). To capture more intricate, nonlinear relationships within the high-dimensional feature set, a Deep Neural Network architecture was developed. The proposed DNN consists of multiple fully connected (dense) layers, utilising nonlinear activation functions to expand its representational capacity (Kim and Kwak 2024). To ensure stable convergence and prevent overfitting, regularisation techniques, including dropout layers and early stopping criteria were integrated into the training process (Ullah *et al.* 2023). For model validation, the dataset was partitioned into training, validation, and testing subsets. Classification performance was rigorously evaluated using a suite of metrics, including accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). This multi-metric approach allowed for a nuanced assessment of model performance, particularly in terms of class-specific discriminative power. Consequently, this comparative framework provides a comprehensive understanding of how different algorithmic structures adapt to the challenges of speech emotion recognition.

Deep Neural Network (DNN) Architecture and Implementation

Deep Neural Networks are robust classification models capable of learning intricate nonlinear relationships through their multi-layered architectures. These models process input data through multiple hidden layers, generating increasingly abstract and discriminative feature representations at each stage, which allows for the successful modelling of complex data patterns (Fayek *et al.* 2017). In a DNN architecture, each layer processes the output from the preceding layer to produce a higher-level representation, thereby enabling hierarchical learning. This characteristic offers a significant advantage, particularly in problems involving high-

dimensional and nonlinear relationships. In speaker-independent emotion recognition tasks, features extracted from acoustic signals are typically high-dimensional and exhibit complex structures. Consequently, DNN-based approaches have been widely adopted in recent years due to their superior ability to learn nonlinear dependencies in such data (Brahmi *et al.* 2024). Particularly in studies using time-frequency representations such as mel-spectrograms, deep learning models have demonstrated an enhanced capacity to capture emotional patterns (Begazo *et al.* 2024). In the present study, the extracted multidimensional acoustic features were utilised as inputs to the DNN model, allowing it to learn the complex interrelationships among these variables (Dixit *et al.* 2024). Within a neural network, each neuron applies a linear transformation by weighting the input data, then generates a nonlinear output via an activation function. This process is formally expressed as follows:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{1}$$

$$\mathbf{a}^{(l)} = \sigma\left(\mathbf{z}^{(l)}\right) \tag{2}$$

During the training process, the loss function was minimised using the backpropagation algorithm, with weights updated iteratively (Fayek *et al.* 2017). To mitigate the risk of overfitting and ensure the model's generalisation capability, regularisation techniques, including dropout, early stopping, and the utilisation of a dedicated validation dataset were strictly employed. This rigorous approach ensures that the model maintains high performance not only on the training data but also on previously unseen test samples (Kim and Kwak 2024).

Specifically, the proposed Deep Neural Network (DNN) architecture consisted of six fully connected hidden layers with 2048, 1024, 512, 256, 128, and 64 neurons, respectively. To mitigate the vanishing gradient problem and ensure smooth nonlinear mapping, the Swish activation function was applied to all hidden layers. Batch Normalization was employed after each dense layer to stabilize the internal covariate shift and accelerate convergence, whereas Dropout layers with rates ranging from 0.20 to 0.45 were strategically inserted to reduce overfitting.

The output layer consisted of eight neurons with a Softmax activation function, producing a probability distribution over the eight emotion classes. Model optimization was performed using the Adam optimizer with an initial learning rate of 5×10^{-4} and the Categorical Cross-Entropy loss function. A label smoothing factor of 0.05 was applied to improve generalization. Training was conducted for a maximum of 600 epochs using a mini-batch size of 64. An EarlyStopping callback with a patience of 50 epochs was employed to restore the model weights corresponding to the lowest validation loss. In addition, the learning rate was automatically reduced by a factor of 0.5 using the ReduceLROnPlateau scheduler whenever the validation loss plateaued, with a minimum learning rate of 1×10^{-6}.

Support Vector Machine: The Support Vector Machine is a robust supervised learning algorithm that identifies the optimal decision boundary (hyperplane) that maximises the margin between classes. It is extensively utilised in speech emotion recognition tasks due to its effectiveness in handling high-dimensional datasets (Cortes and Vapnik 1995). The primary objective of SVM is to minimise the norm of the weight vector, which is equivalent to maximising the margin between the support vectors of different classes. This optimisation problem is formally defined as:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3}$$

subject to the following constraint for all training samples:

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \qquad i = 1, \ldots, N \tag{4}$$

where $\mathbf{w} \in \mathbb{R}^d$ represents the weight vector (normal vector) that defines the orientation of the optimal hyperplane, $b \in \mathbb{R}$ denotes the bias (intercept) term, which determines the offset of the hyperplane from the origin, $\mathbf{w}^T\mathbf{x}_i + b$ represents the linear decision function, where the superscript T denotes the transpose operator, ensuring a proper dot product between vectors, and $\|\mathbf{w}\|$ represents the Euclidean norm of the weight vector.

However, since emotional acoustic features are rarely linearly separable in real-world datasets, kernel functions are employed to model nonlinear decision boundaries. In this study, the Radial Basis Function (RBF) kernel was utilised to map data points into a higher-dimensional space, enabling the SVM to effectively capture complex relationships and distinguish between emotion categories with high precision.

In this study, the Support Vector Machine (SVM) classifier was configured with a Radial Basis Function (RBF) kernel to effectively handle the nonlinear boundaries within the high-dimensional speech feature space. The regularization parameter was set to $C = 10$ to achieve an optimal balance between maximizing the decision margin and minimizing training errors. The kernel coefficient (γ) was dynamically computed using the "scale" heuristic, defined as

$$\gamma = \frac{1}{n_{\text{features}} \times \text{Var}(X)} \tag{5}$$

to adapt to the variance of the extracted acoustic features. Furthermore, to mitigate the influence of class imbalance and prevent the classifier from biasing toward dominant classes, cost-sensitive learning was incorporated by applying class weights inversely proportional to class frequencies during model training.

Random Forest: Random Forest is a robust ensemble learning method that integrates multiple decision trees based on the bootstrap aggregating (bagging) principle. In this approach, each decision tree is trained on independent bootstrap samples of the dataset. By exposing each tree to different data distributions, the model's overall generalisation capability is significantly enhanced. Furthermore, at each node during the splitting process, a randomly selected subset of features is utilised rather than the entire feature set. This mechanism reduces the correlation between individual trees, effectively mitigating the risk of overfitting. While individual decision trees classify data by hierarchically partitioning the feature space, using criteria such as Information Gain or Gini Impurity, they are often prone to high variance and overfitting. The Random Forest architecture addresses these limitations by aggregating the outputs of numerous trees to yield more stable and robust predictions. Consequently, RF models exhibit superior performance, particularly in high-dimensional feature spaces and when dealing with noisy data (Jha *et al.* 2022).

The final classification result in a Random Forest is determined by majority voting, combining the predictions from all constituent trees. This process is formally defined as:

$$\hat{\mathbf{y}} = \text{mode}\left(T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_n(\mathbf{x})\right) \tag{6}$$

where $\mathbf{x}$ represents the input feature vector, $T_i(\mathbf{x})$ denotes the class prediction generated by the i-th decision tree in the forest, n is the total number of decision trees (bootstrap samples) in the forest, and $\hat{\mathbf{y}}$ represents the final predicted class label.

A significant advantage of the Random Forest model is its inherent ability to calculate feature importance, which identifies the acoustic features that are most discriminative for emotion classification. This interpretability is particularly valuable in speech emotion recognition, where understanding the contribution of specific spectral and temporal features is essential. Consistent with the literature, the Random Forest algorithm was selected because of its reliability and stable performance on complex, high-dimensional datasets24.

For the ensemble-based evaluation, the Random Forest (RF) classifier was implemented using 800 decision trees ($n_{\text{estimators}} = 800$) to ensure prediction stability and reduce variance. To address class imbalance during tree construction, the `balanced_subsample` strategy was employed, which automatically computes class weights from the bootstrap sample at each tree. Finally, to ensure reproducibility and enable consistent comparisons across different experimental runs, the pseudo-random number generator was initialized with a fixed seed (random_state = 42).

Evalutation Metrics

In this study, a comprehensive set of evaluation metrics, including accuracy, precision, recall, and F1-score was utilised to assess the performance of the speech-based emotion recognition models. These metrics facilitate a detailed analysis of both the overall system efficacy and class-specific performance (Jha *et al.* 2022). Accuracy represents the ratio of correct predictions to the total number of samples. While it provides a preliminary assessment of overall performance, it is often insufficient for evaluating multi-class tasks with imbalanced datasets. Consequently, Precision, Recall (Sensitivity), and F1-score were integrated to examine class-specific discriminative power (Serrano *et al.* 2025). Precision measures the reliability of the model's positive predictions, whereas recall evaluates the model's ability to identify all relevant instances within a specific emotional category correctly. To provide a balanced evaluation of these two metrics, the F1-score, the harmonic mean of precision and recall was employed as a more robust indicator of performance under class-imbalance conditions (Shah *et al.* 2023).

For the multi-class emotion recognition problem, both macro and weighted averages were calculated. The macro average assigns equal importance to all emotion classes regardless of their frequency, whereas the weighted average adjusts each class's contribution based on its sample size in the dataset. Furthermore, the Receiver Operating Characteristic curves and the Area Under the Curve metric were utilised to evaluate the models' discriminative performance (Brahmi *et al.* 2024). The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different threshold values. At the same time, the AUC quantitatively reflects the model's ability to distinguish a target emotion from other classes. In this multi-class framework, ROC-AUC values were calculated for each emotion category individually, providing a granular analysis of the model's strengths and limitations across different emotional states.

Experimental Results

In the scope of this research, the performance of the Support Vector Machine (SVM), Random Forest (RF), and Deep Neural Network (DNN) models was systematically evaluated. The comparative analysis indicates that each architecture exhibited distinct behaviours depending on the acoustic complexity of the emotion classes.

The learning dynamics of the DNN model, as depicted in the accuracy and loss curves in Figure 4, demonstrated a highly efficient optimisation process. Although the model achieved a training accuracy of 97%, the validation accuracy stabilised at 87%, highlighting a trade-off between learning capacity and generalisation. In contrast, the SVM model achieved the highest overall classification accuracy of 79.51%, indicating that its margin-maximisation principle is particularly effective for high-dimensional acoustic feature vectors used in speech emotion recognition. The Random Forest model achieved an overall accuracy of 71.18%. Despite its ensemble-based structure, its performance was relatively lower, which may be attributed to its greater sensitivity to noise present in prosodic features.

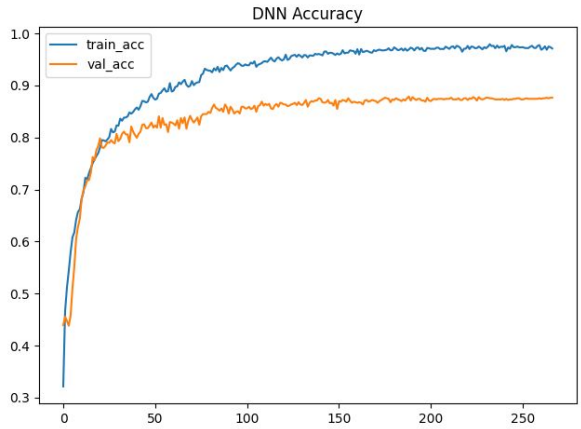


Figure 4 Training and validation performance of the DNN model

The detailed classification performance of the DNN model was further investigated using the confusion matrix (Figure 5) and the comparative performance chart (Figure 6). The model, evaluated on 1,152 test samples, achieved varying levels of performance across the emotion categories. As illustrated in Figure 6, the highest recognition performance was obtained for the Angry (0.85) and Disgust (0.81) classes. From a mathematical perspective, the high precision and recall achieved for these categories indicate that their acoustic signatures, characterised by elevated energy levels and pronounced frequency variations, occupy relatively distinct regions within the feature space.

In contrast, the Sad class presented the greatest challenge for the DNN model, yielding an F1-score of 0.59. The confusion matrix demonstrates that sad speech samples were frequently misclassified as Calm or Neutral, primarily because of the substantial overlap among low-arousal emotional speech patterns. A notable observation was obtained for the Neutral class, where a high recall (0.81) was accompanied by a comparatively low precision (0.50). This result indicates a false-positive tendency, suggesting that the model frequently assigned ambiguous low-energy speech signals to the Neutral category. Furthermore, this behaviour was intensified by the class imbalance present in the dataset, which biased the decision boundaries toward classes with a larger number of training samples and consequently reduced the classification performance for the underrepresented Neutral class.

The discriminative performance of the models was further evaluated using Receiver Operating Characteristic (ROC) analysis. SVM model demonstrated consistently superior performance, with Area Under the Curve (AUC) values ranging from 0.98 to 0.99 across all emotion classes. These high AUC values indicate an excellent ability to discriminate between positive and negative classes over a wide range of decision thresholds.

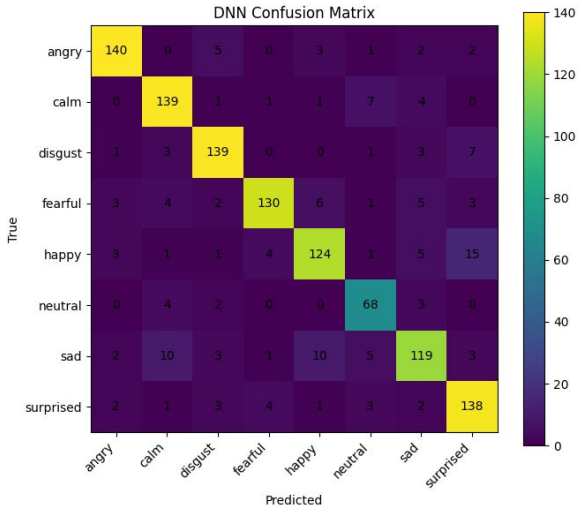


Figure 5 Confusion matrix of the DNN model for speech emotion recognition across eight emotional categories

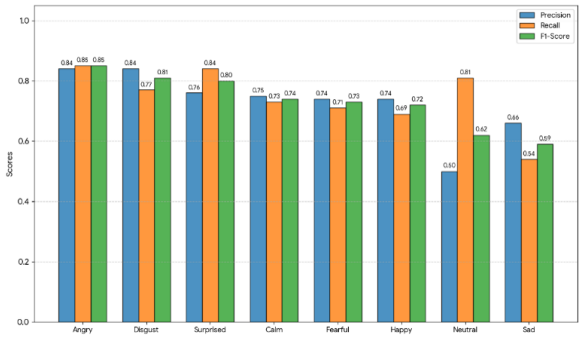


Figure 6 Comparative analysis of class-specific precision, recall, and F1-score for the DNN model

For the Random Forest model, although the Calm and Disgust classes achieved AUC values of 0.97, the AUC for the Sad class decreased to 0.91, indicating a reduced ability to distinguish acoustically similar emotional categories. The DNN model also achieved high overall AUC values; however, its discriminative performance for the Happy and Surprised classes was comparatively lower because of the similar high-arousal vocal characteristics shared by these emotions, resulting in increased inter-class confusion.

Overall, the experimental findings indicate that the SVM model provides the most consistent and balanced performance for speech emotion recognition. Although the DNN exhibited substantial learning capacity, its lower validation performance suggests that additional data augmentation or regularization strategies may be required to further improve generalization. The Random Forest model remained a competitive alternative but achieved lower overall performance for the selected acoustic feature set.

Ultimately, the results confirm that inter-class acoustic similarities, particularly between the Happy–Surprised and Sad–Calm emotion pairs, represent the primary limitation in achieving perfect classification accuracy. These findings suggest that the selection of an appropriate classification model should be guided by the acoustic characteristics and distribution of the target emotion classes to develop robust and reliable speech emotion recognition systems.

CONCLUSION

In this study, a comprehensive comparative evaluation was conducted to assess the performance of various machine learning and deep neural network architectures in speech emotion recognition. Specifically, SVM, RF, and DNN models were rigorously trained and analysed using a multidimensional feature set extracted from time, frequency, and cepstral domains. The experimental findings indicate that features derived from Mel-Frequency Cepstral Coefficients and Mel-spectrograms play a quintessential role in determining classification accuracy. The synergistic use of these spectral and cepstral representations allowed the models to more effectively characterise the complex interplay between the temporal dynamics and frequency components of emotional speech signals.

Regarding model performance, the SVM model achieved the best overall accuracy and F1-score. While the DNN model exhibited strong predictive performance, it showed limited generalisation, particularly when validation set results were compared with those from the training phase. In contrast, the Random Forest model consistently underperformed its counterparts, struggling to capture the non-linear relationships in the acoustic feature space. A granular analysis of the ROC curves on a class-by-class basis further validates these conclusions. The SVM model achieved exceptionally high and balanced Area Under the Curve values across all emotional categories, representing a robust discriminative threshold. While the DNN model also demonstrated substantial discriminative performance, it was notably lower for the happy and sad classes than for other emotional states. This performance degradation in specific categories can be theoretically attributed to the inherent acoustic similarities and overlapping pitch/energy profiles shared by these emotions. Furthermore, the simultaneous evaluation of the DNN model's training and validation metrics revealed a significant performance gap: although the model achieved near-perfect accuracy on the training data, it failed to maintain this level during validation, indicating a persistent overfitting problem.

This finding underscores that the model's architecture, while high in learning capacity, requires further refinement to improve generalisation to unseen data. Detailed confusion matrix analyses further corroborated this, showing a heightened frequency of confusion between happy-surprised and sad-calm pairs. These misclassifications suggest that certain emotional states remain difficult to isolate due to their closely related acoustic manifestations in the vocal signal. In conclusion, the strategic combination of diverse acoustic features has been identified as a critical factor in optimising speech emotion recognition systems. The study highlights that while the SVM model stands out for its robustness, the DNN model remains a high-capacity tool that requires advanced regularisation techniques, data augmentation, and more balanced datasets to overcome its current generalisation constraints. Future research will focus on integrating more sophisticated deep learning architectures and novel feature combinations to further refine the sensitivity and specificity of emotion recognition from speech (Kim and Kwak 2024).

Availability of data and material

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Akçay, M. B. and K. Oğuz, 2020 Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication **116**: 56–76.

Begazo, R., A. Aguilera, I. Dongo, and Y. Cardinale, 2024 A combined cnn architecture for speech emotion recognition. Sensors **24**.

Brahmi, Z., M. Mahyoob, M. Al-Sarem, J. Algaraady, K. Bousselmi, *et al.*, 2024 Exploring the role of machine learning in diagnosing and treating speech disorders: A systematic literature review. Psychology Research and Behaviour Management **17**: 2205–2232.

Çolakoğlu, E., S. Hızlısoy, and R. S. Arslan, 2022 Konuşmadan duygu tanıma üzerine detaylı bir inceleme: Özellikler ve sınıflandırma metotları. European Journal of Science and Technology .

Cortes, C. and V. Vapnik, 1995 Support-vector networks. Machine Learning **20**: 273–297.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, *et al.*, 2001 Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18**: 32–80.

Cui, W., F. Jiang, X. Gao, S. Zhang, and D. Zhao, 2018 An efficient deep quantized compressed sensing coding framework of natural images. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1777–1785.

Dixit, S., D. M. Low, G. Elbanna, F. Catania, and S. S. Ghosh, 2024 Explaining deep learning embeddings for speech emotion recognition by predicting interpretable acoustic features. arXiv preprint arXiv:2409.09511 .

Durukal, M. and A. K. Hocaoglu, 2015 Performance analysis of mfcc features on emotion recognition from speech. International Journal of Scientific and Technological Research **1**.

Fayek, H. M., M. Lech, and L. Cavedon, 2017 Evaluating deep learning architectures for speech emotion recognition. Neural Networks **92**: 60–68.

Gales, M. J. F., K. M. Knill, and S. J. Young, 1999 State-based gaussian selection in large vocabulary continuous speech recognition using hmms. IEEE Transactions on Speech and Audio Processing **7**: 152–161.

Hsu, C.-C., J. Krajewski, and J. Felfe, 2024 How speech prosody affects leadership perceptions: A machine learning approach. International Journal of Organizational Leadership **13**: 432–450.

Jha, T., R. Kavya, J. Christopher, and V. Arunachalam, 2022 Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. International Journal of Speech Technology **25**: 707–725.

Jiang, X. and M. D. Pell, 2018 Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments. In *Proceedings of the International Conference on Speech Prosody*, pp. 269–273.

Juslin, P. N. and P. Laukka, 2003 Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin **129**: 770–814.

Kane, J., M. N. Johnstone, and P. Szewczyk, 2024 Voice synthesis improvement by machine learning of natural prosody. Sensors **24**: 1624.

Kim, T. W. and K. C. Kwak, 2024 Speech emotion recognition using deep learning transfer models and explainable techniques. Applied Sciences **14**.

Livingstone, S. R. and F. A. Russo, 2018 The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PLOS ONE **13**: e0196391.

McLaren, M., A. Lawson, Y. Lei, and N. Scheffer, 2013 Adaptive gaussian backend for robust language identification. In *Interspeech 2013*, pp. 84–88.

Ming, Y., B. Lyu, and Z. Li, 2023 Cast: Context-association architecture with simulated long-utterance training for mandarin speech recognition. Speech Communication **155**: 102985.

Mustaqeem and S. Kwon, 2019 A cnn-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**: 183.

Pala, M. A., 2025 Specedgenet: A hybrid graph neural network architecture integrating information from spectrum to bond for drug property prediction. In *Proceedings of the 2025 Innovations in Intelligent Systems and Applications Conference (ASYU 2025)*.

Serrano, S., O. Serghini, G. Esposito, S. Carbone, C. Mento, *et al.*, 2025 Review and comparative analysis of databases for speech emotion recognition. Data **10**.

Shah, N., K. Sood, and J. Arora, 2023 Speech emotion recognition for psychotherapy: an analysis of traditional machine learning and deep learning techniques. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 718–723.

Tao, J. and T. Tan, 2005 Affective computing: A review. In *International Conference on Affective Computing and Intelligent Interaction*.

Ullah, R., M. Asif, W. A. Shah, F. Anjam, I. Ullah, *et al.*, 2023 Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. Sensors **23**: 6212.

Wu, C.-H., J.-F. Yeh, Z.-J. Chuang, and C. Yi, 2002 Emotion perception and recognition from speech. LDC Catalog.

Yılmazcan, D. S. and M. A. Pala, 2026 Exploring the chemical space of bace-1 inhibitors: Structure-based prediction with deep learning and machine learning. Computers and Electronics in Medicine **3**: 36–41.

How to cite this article: Baydar, S. E., and Pala M. A. Efficacy of Multi-Domain Acoustic Features in Speech Emotion Recognition: A Comparative Analysis of Machine Learning and Deep Learning Models. *Chaos and Fractals*, 3(2), 117-124, 2026.

ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
REVIEW ARTICLE
Vol.3 / No.2 / 2026 / pp.125-136
https://doi.org/10.69882/adba.chf.2026076

From Deterministic Chaos to Machine Learning: A Comparative Review of Pseudo-Random Number Generators for Modern Computing and Cryptography

Mutala Abdul Karim [*,1], **Peter Awon-natemi Agbedemnab** [β,2] **and Salamudeen Alhassan** [β,3]

[*]Department of Mathematics, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana, [β]Department of Information Systems and Technology, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana.

ABSTRACT Pseudo-random number generators are central to simulation, cryptography, embedded systems, privacy-preserving computation, and machine-learning workflows. This review compares traditional, chaos-based, hybrid, residue-number-system-based, and machine-learning-assisted approaches using computational efficiency, period length, entropy, autocorrelation, statistical-test performance, scalability, and cryptographic suitability. Traditional generators such as linear congruential generators, linear feedback shift registers, Xorshift, and the Mersenne Twister remain attractive for high-throughput simulation because they are simple, reproducible, and fast. Their linearity and recoverable internal states, however, limit their suitability for security-sensitive applications. Chaos-based generators improve nonlinearity, sensitivity to initial conditions, and entropy, but their performance depends strongly on parameter tuning and finite-precision implementation. Hybrid and residue-number-system designs offer a promising middle ground by combining modular arithmetic, nonlinear dynamics, and parallel residue computation. Machine-learning techniques further expand generator evaluation and design by detecting hidden bias, learning complex output patterns, and supporting adaptive generation, although they introduce training cost, reproducibility concerns, and hardware requirements. Future research should prioritize secure hybrid architectures, precision-aware chaotic maps, adaptive reseeding, hardware acceleration, and standardized testing frameworks for nonlinear and data-driven randomness systems.

KEYWORDS
Pseudo-random number generators
Cryptography
Chaotic maps
Machine learning
Hybrid generators
Randomness testing
Residue number system
Security

INTRODUCTION

Random number generation is a foundational requirement in modern computing. It supports simulation, statistical sampling, secure key generation, cryptographic protocols, randomized algorithms, gaming, embedded security, and machine-learning workflows (Maheshwari *et al.* 2014; Noibate 2023). In practice, random number generators are commonly grouped into true random number generators (TRNGs), which derive uncertainty from physical processes, and pseudo-random number generators (PRNGs), which produce deterministic sequences from mathematical rules and initial seeds.

PRNGs are especially important because they are fast, reproducible, and easy to implement in software or hardware. Their deterministic structure allows the same seed to reproduce the same sequence, which is useful in simulation and experimental repeatability (Maheshwari *et al.* 2014; Martini and Bruno 2017). This same determinism creates security risks when internal states, recurrence rules, or seeds can be inferred from observed outputs. Therefore, a useful PRNG must balance statistical randomness, computational cost, scalability, reproducibility, and resistance to prediction.

This review compares five major families of PRNG-related methods: traditional arithmetic and bitwise generators, cryptographically secure PRNGs, chaos-based generators, hybrid/residue-number-system designs, and machine-learning-assisted approaches. Rather than treating description and comparison as separate stages, Section embeds a short critical synthesis immediately after each family is introduced, so that similarities, differences, advantages, and limitations are visible where the algorithms are first discussed. Section consolidates these observations into an explicit, criterion-by-criterion comparative analysis, Section discusses the resulting trade-offs and future directions, and Section concludes the review.

Manuscript received: 16 June 2026,
Revised: 14 July 2026,
Accepted: 14 July 2026.

[1]mutalaak@gmail.com (**Corresponding author**)
[2]pagbedemnab@cktutas.edu.gh
[3]salhassan@cktutas.edu.gh

The evaluation criteria used throughout are period length, entropy, autocorrelation, NIST/Diehard test performance, throughput, memory requirement, hardware suitability, and cryptographic applicability. The central argument is that no single generator is universally optimal; the strongest design depends on the target application and the acceptable trade-off between speed, security, and implementation complexity.

Figure 2 presents a taxonomy of all PRNG families discussed in this review, and it is referenced throughout Section as each branch is introduced in turn.

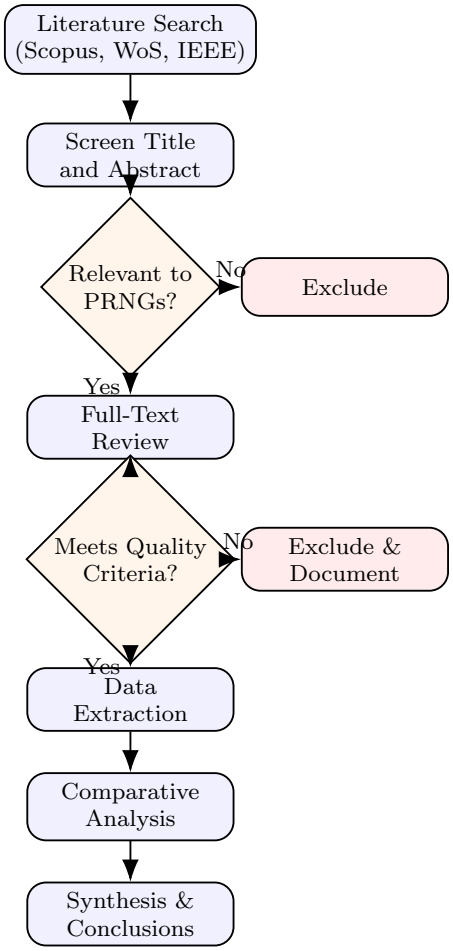


Figure 1 Review methodology flowchart. Articles retrieved from Scopus, Web of Science, and IEEE Xplore were screened by title and abstract, assessed for relevance, quality-filtered, then data-extracted for comparative analysis

The literature underlying this review was retrieved from Scopus, Web of Science, and IEEE Xplore, and was screened in the four stages summarized in Figure 1: an initial title/abstract screen for relevance to PRNGs, a full-text review, a quality-criteria filter, and a final data-extraction step feeding the comparative analysis and synthesis presented in Sections and . Records that failed the relevance screen or the quality filter were excluded and, where applicable, the reason for exclusion was documented. This staged process produced the reference set summarized in Tables 4 and 5 and discussed in Section .

LITERATURE REVIEW

Traditional PRNG Approaches

Traditional PRNGs, shown on the left branch of Figure 2, were developed mainly for numerical simulation, Monte Carlo analysis, and general-purpose computation. They rely on fixed recurrence relations, modular arithmetic, bitwise transformations, or finite-field operations rather than adaptive learning. Their main advantages are simplicity, speed, and low memory cost, while their main weaknesses are predictability and limited resistance to state-recovery attacks (Noibate 2023; Saeed *et al.* 2024).

Linear Congruential Generator (LCG): The LCG is a simple and popular approach to producing a string of pseudo-random numbers, defined by

$$X_{n+1} = (aX_n + c) \bmod m. \tag{1}$$

The series depends on an initial seed X_0; the Hull–Dobell theorem establishes conditions for achieving the longest period m, and Marsaglia showed that LCG outputs lie on a finite number of parallel hyperplanes, a structural weakness exploited in state-recovery attacks (Martini and Bruno 2017; Noibate 2023).

Quadratic Congruential Generator (QCG): The QCG is a nonlinear extension of the LCG,

$$X_{n+1} = (aX_n^2 + bX_n + c) \bmod m, \tag{2}$$

where $m > 0$ is the modulus and a, b, and c are integer parameters (Noibate 2023; Sayed *et al.* 2016).

Lagged Fibonacci Generator (LFG):

$$X_n = (X_{n-j} \oplus X_{n-k}) \bmod m, \tag{3}$$

with $0 < j < k$ integer lags and the initial k values forming the seed vector.

Mersenne Twister (MT19937): Developed by Matsumoto and Nishimura in 1998, MT19937 produces 32-bit integers with period $2^{19937} - 1$ (Martini and Bruno 2017; Oliveira 2024), with recurrence over GF(2):

$$X_{n+1} = f(X_n, X_{n+1}, \ldots, X_{n+k}). \tag{4}$$

Wichmann–Hill Combined LCG (CLCG):

$$X_n = \left(\sum_{i=1}^{r} a_i X_n \bmod m_i \right) \bmod 1, \tag{5}$$

$$U_n = \left(\sum_{i=1}^{r} \frac{X_{n,i}}{m_i} \right) \bmod 1, \tag{6}$$

where r is the number of combined generators, a_i the multiplier, and m_i the modulus of generator i.

Middle Square Method: One of the earliest pseudo-random methods, it squares a four-digit seed to produce an eight-digit number and uses the middle four digits as the next seed. It has a very short period ($< 10^4$) and is unsuitable for large-scale simulation (Martini and Bruno 2017; Saeed *et al.* 2024).

Xorshift Generator:

$$X_i = X_{i-1} \oplus (X_{i-1} \ll a) \oplus (X_{i-1} \gg b) \oplus (X_{i-1} \ll c), \tag{7}$$

with $\oplus$ denoting bitwise exclusive OR and $\ll$, $\gg$ denoting left and right bit-shifts.

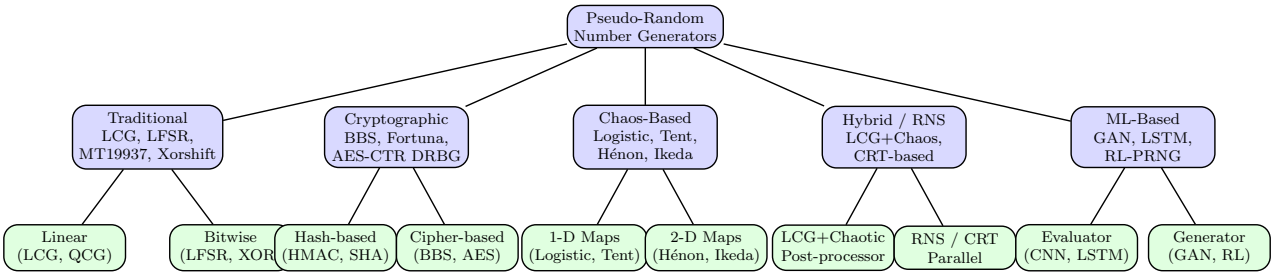


Figure 2 Taxonomy of PRNG families and sub-types. Five major families branch from the root: traditional, cryptographic, chaos-based, hybrid/residue-number-system, and machine-learning-based. Each family is further decomposed into its principal sub-types. Machine-learning-based generators serve both generative and evaluative roles in PRNG research

Linear Feedback Shift Register (LFSR):

$$b_i = c_1 b_{i-1} \oplus c_2 b_{i-2} \oplus \cdots \oplus c_N b_{i-N}, \tag{8}$$

with binary tap coefficients $c_1, \ldots, c_N \in \{0, 1\}$.

Synthesis: Table 1 consolidates the recurrence, period, speed, and security profile of every traditional generator discussed above. Throughput is highest exactly where the recurrence is linear and bitwise, as in Xorshift and LFSR, and this same linearity is what a Berlekamp–Massey-style analysis or hyperplane analysis exploits to recover the internal state. None of the generators in this family should therefore be used unmodified for key generation or nonce derivation. MT19937 stands out for its exceptionally long period and strong empirical statistics, but its state is fully recoverable from 624 consecutive 32-bit outputs, so its suitability is restricted to simulation and non-adversarial sampling. This trade-off motivates the cryptographic designs discussed next.

Cryptographically Secure PRNGs

Where traditional generators fail is exactly where cryptographically secure PRNGs are designed to succeed: unpredictability under adversarial observation.

Blum–Blum–Shub (BBS): Proposed in 1986, BBS is founded on the hardness of integer factorization, the same hard problem underlying RSA:

$$X_{n+1} = X_n^2 \bmod M, \qquad M = pq, \tag{9}$$

with p and q large safe primes.

Fortuna: Described by Ferguson and Schneier in 2003, Fortuna improves on Yarrow by addressing entropy-estimation flaws and providing perpetual reseeding from multiple entropy sources using AES-based entropy-accumulation pools.

Yarrow: Invented by Schneier, Kelsey, and Ferguson in the late 1990s, Yarrow applies cryptographic post-processing through block ciphers and hash functions to produce sequences that remain unpredictable in a hostile environment (Almaraz Luengo and Román Villaizán 2023).

ANSI X9.17 (1985) / X9.31 (1998): This family uses block ciphers such as Triple-DES or AES in feedback mode (Oliveira 2024):

$$X_n = E_k(DT) \oplus R_{n-1}, \tag{10}$$

where DT is a date/time value and E_k is encryption under key k.

AES-CTR DRBG, HMAC-DRBG, and SHA-256 Hash DRBG: These modern deterministic random-bit generators are used throughout current cryptographic systems (Guisande *et al.* 2023). AES-CTR DRBG encrypts successive counter values,

$$R_i = E_k(V + i), \tag{11}$$

HMAC-DRBG replaces encryption with a keyed hash,

$$V_i = \mathrm{HMAC}_k(V_{i-1}), \quad K = \mathrm{HMAC}_k(V_i \| \mathrm{entropy}), \tag{12}$$

and Hash-DRBG applies a cryptographic hash to a seed-derived state,

$$R_i = \mathrm{Hash}(V + i), \quad V \leftarrow \mathrm{Hash}(V \| \mathrm{const} \| \mathrm{entropy}). \tag{13}$$

Synthesis: As the cryptographic rows of Table 1 show, every cryptographically secure design trades throughput for security. Each relies on either a hard number-theoretic problem, a block cipher, or a keyed hash precisely to remove the linear recoverability that limits traditional generators, at the cost of lower speed.

Background: TRNGs and the Physical Basis of Randomness

True random numbers derive from unpredictable physical processes rather than deterministic algorithms; unlike PRNGs, true randomness comes from stochastic natural phenomena (Martini and Bruno 2017). Table 2 summarizes the principal TRNG source types, their entropy rates, and typical applications.

Table 2 shows a clear entropy ordering: quantum-mechanical sources such as photon detection, quantum tunnelling, and radioactive decay reach the highest entropy classification because their underlying physical models are fundamentally stochastic rather than merely complex. Classical noise sources and oscillator/metastability-based sources are rated lower because residual correlations or environmental coupling can, in principle, be modeled. Environmental sources such as keystrokes and mouse movement are the weakest and are typically used only to seed a PRNG or cryptographically secure generator. This entropy hierarchy provides the physical justification for chaos-based and hybrid PRNGs, which attempt to approximate high-entropy behavior algorithmically when a physical source is unavailable or too slow.

Chaos-Based PRNGs

Chaos maps generate complex, seemingly random behavior from deterministic systems. Despite being governed by simple equations, they exhibit sensitivity to initial conditions, the "butterfly effect" making them useful in cryptography, random number generation, and nonlinear-dynamics research (Saeed *et al.* 2024).

Table 1 Comparative study of traditional and related pseudo-random number generator approaches

PRNG type	Application	Method	Equation/operation	Period	Speed	Security
LCG	Simulation	Modular linear	$X_{n+1} = (aX_n + c) \bmod m$	Up to m	Fast	Low
QCG	Simulation	Quadratic modular	$X_{n+1} = (aX_n^2 + bX_n + c) \bmod m$	Up to m	Moderate	Low
LFG	Monte Carlo	Recurrence	$X_n = (X_{n-j} \oplus X_{n-k}) \bmod m$	$> 10^{17}$	Moderate	Low
MT19937	Scientific	Recurrence over GF(2)	$X_{n+1} = f(X_n, \ldots, X_{n+k})$	$2^{19937} - 1$	Fast	Low
Wichmann–Hill	Simulation	Multi-modular	$U_n = (\sum X_{n,i}/m_i) \bmod 1$	$> 10^{18}$	Moderate	Low
Middle Square	Historical	Digit extraction	Middle digits of X_n^2	$< 10^4$	Slow	Very low
Xorshift	Simulation, GPU	XOR and shift	$X_i = X_{i-1} \oplus (\cdots)$	$2^{128} - 1+$	Very fast	Low
LFSR	Hardware, communications	Shift and feedback	$b_i = c_1 b_{i-1} \oplus \cdots$	$2^n - 1$	Very fast	Low
BBS	Cryptography	Modular squaring	$X_{n+1} = X_n^2 \bmod M$	$\sim 2^n$	Slow	Very high
Fortuna	Cryptography	AES and entropy	AES-based reseeding	Extremely long	Moderate	Very high
Yarrow	Cryptography	Entropy and reseeding	Block-cipher pools	Extremely long	Moderate	Very high
ANSI X9.17/31	Financial	AES and timestamp	$X_n = E_k(DT) \oplus R_{n-1}$	Extremely long	Moderate	Very high
AES-CTR DRBG	Cryptography	Hash/cipher	$R_i = E_k(V + i)$	Extremely long	Moderate	Very high
Chaotic PRNGs	Cryptography, research	Chaotic dynamics	$x_{n+1} = rx_n(1 - x_n)$	Parameter-dependent	Moderate	High
Hybrid PRNGs	Cryptography, simulation	Arithmetic and chaos	LCG $\oplus$ chaotic map	Very long	Moderate	High
RNS-based	Cryptography, HPC	Multi-modular	CRT reconstruction	Very long	Moderate	High

Table 2 Comparative overview of true random number generator sources. Entropy rate is relative to the class of generators considered

Type	Randomness source	Generation method	Mathematical model	Entropy	Applications
Thermal noise	Electron thermal agitation	Digitize analog voltage noise	$V_n^2 = 4kTR\Delta f$	High	Hardware RNGs; secure chips
Shot noise	Discrete electron flow	Current fluctuation detection	$I_n^2 = 2qI\Delta f$	High	Embedded security hardware
Avalanche noise	Reverse-biased diode breakdown	Amplify and digitize diode noise	$I_n = \sqrt{2qI_a\Delta f}\, G$	Very high	TPM modules; crypto hardware
Photonic TRNG	Photon detection/quantum uncertainty	Arrival times or polarization	$P(n) = \lambda^n e^{-\lambda}/n!$	Extremely high	Quantum cryptography
Quantum tunnelling	Electron tunnelling through barriers	Tunnelling-current measurement	$T = e^{-2\kappa d}$	Extremely high	Quantum key generation
Radioactive decay	Nuclear decay events	Inter-arrival-time sensors	$P(t) = \lambda e^{-\lambda t}$	Extremely high	Scientific randomness
Oscillator jitter	Clock phase noise	Jitter between asynchronous oscillators	$\Delta t_n = \phi_n/(2\pi f)$	Moderate	Microcontrollers; FPGAs
Metastability	Unstable flip-flop states	Latch uncertain state and sample	$\tau = \tau_0 e^{\Delta V/V_t}$	Moderate	FPGA; IoT devices
SRAM startup	Initial memory-cell states	Read startup bits after power-on	$P(1) = \frac{1}{2} + \delta, \delta \to 0$	Moderate	Embedded security; IoT
Environmental	Keystrokes, mouse, weather	Sample real-world events	$H = -\sum p_i \log_2 p_i$	Low–moderate	Seeding; monitoring; simulation

Representative Chaotic Maps

Tent Map:

$$x_{n+1} = \begin{cases} \mu x_n, & x_n < \frac{1}{2}, \\ \mu(1 - x_n), & x_n \geq \frac{1}{2}, \end{cases} \qquad \mu \in [0, 2]. \tag{14}$$

(Sayed *et al.* 2016).

Ikeda Map:

$$x_{n+1} = 1 + \mu(x_n \cos t_n - y_n \sin t_n), \tag{15}$$
$$y_{n+1} = \mu(x_n \sin t_n + y_n \cos t_n), \tag{16}$$

where $t_n = 0.4 - 6/(1 + x_n^2 + y_n^2)$ and $\mu \approx 0.9$ (Oliveira 2024).

Logistic Map:

$$x_{n+1} = rx_n(1 - x_n), \qquad x_n \in [0, 1], \quad r \in [0, 4]. \tag{17}$$

For $r > 3.57$ the map exhibits deterministic chaos (Aniszewska 2018).

Tinkerbell Map:

$$x_{n+1} = x_n^2 - y_n^2 + ax_n + by_n, \tag{18}$$
$$y_{n+1} = 2x_n y_n + cx_n + dy_n, \tag{19}$$

with control parameters $a = 0.9$, $b = -0.6013$, $c = 2$, and $d = 0.5$ (Sathya *et al.* 2021).

Hénon Map:

$$x_{n+1} = 1 - ax_n^2 + y_n, \qquad y_{n+1} = bx_n, \tag{20}$$

with $a = 1.4$ and $b = 0.3$ (Guisande *et al.* 2023).

Chirikov Standard Map:

$$P_{n+1} = P_n + K\sin(\theta_n) \bmod 2\pi, \tag{21}$$
$$\theta_{n+1} = \theta_n + P_{n+1} \bmod 2\pi, \tag{22}$$

where θ_n is the angular position, P_n the conjugate momentum, and K the nonlinearity parameter (Silva and Prado 2023).

Discussion of Table 3: No single map dominates across all criteria. Logistic and Tent maps are preferred where speed and implementation simplicity matter most, but both are known to collapse to short cycles under finite-precision arithmetic and are not considered secure in isolation. Two-dimensional maps such as Hénon, Ikeda, and Tinkerbell trade higher computational cost for stronger, higher-dimensional mixing and are consequently favored in image encryption and hybrid designs. Chirikov and Chebyshev maps offer strong ergodic mixing at the highest floating-point cost. This trade-off between map dimensionality, mixing strength, and computational cost recurs throughout the review. Figure 3 illustrates the transition to chaos for the logistic map, the most widely studied map in PRNG design.

Review of Related Literature on Chaos-Based PRNGs: Al-Daraiseh *et al.* (2023) introduced the State-Based Tent-Map, a modified chaotic random-number generator addressing key limitations of the original Tent map, and reported a 97.4% pass rate across 91,200 Diehard tests. De Micco *et al.* (2021) presented ROSEUR-MOD, which uses modified Euler discretization to convert continuous chaotic systems into discrete maps; it passes all NIST and Diehard tests with reduced precision and is efficient on field-programmable gate arrays. Irfan *et al.* (2020) proposed HC-MRLM with Control of Chaos, achieving a uniform binary distribution and passing all 15 NIST SP 800-22 tests. Wang and Cheng (2019) proposed a PRNG based on the logistic chaotic system, transforming logistic outputs into 15-digit numbers and applying modular reduction to form binary sequences. Yu *et al.* (2019) introduced a PRNG combining a four-wing memristive hyperchaotic system with a Bernoulli map, achieving 62.5 Mbps in an FPGA implementation while passing rigorous statistical tests. Ismail *et al.* (2018) presented a hybrid Tent–Logistic PRNG for secure communications that passes NIST SP 800-22 while remaining computationally lightweight.

Machicao and Bruno (2017) proposed a deep-zoom technique to extract the k rightmost digits of chaotic orbits, competing with the Mersenne Twister in statistical performance. Jiang *et al.* (2017) developed a true random number generator based on stochastic threshold switching in an Ag:SiO_2 diffusive memristor, passing all 15 NIST tests without post-processing. Avaroglu (2017) proposed a PRNG combining two Arnold cat maps, passing NIST tests with a hardware-friendly design. Stoyanov and Kordov (2015) proposed a PRNG based on two Tinkerbell maps with a key space of 2^{183}, passing NIST, DIEHARD, and ENT tests; earlier work combined two Chirikov standard maps with a modified shrinking rule and also passed all three suites (Stoyanov and Kordov 2014).

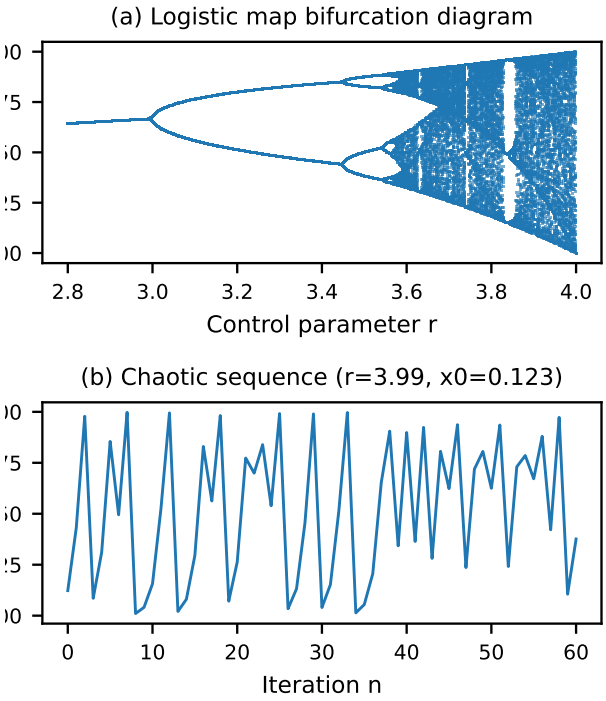


Figure 3 Chaotic behavior of the logistic map in Equation (17). (a) Bifurcation diagram for $r \in [2.8, 4.0]$: period-doubling cascades begin near $r \approx 3.0$, and deterministic chaos emerges near $r \approx 3.57$. (b) A representative chaotic sequence at $r = 3.99$, $x_0 = 0.123$, illustrating aperiodic, noise-like output

Discussion of Tables 4 and 5: Read together, the tables show that essentially every chaos-based design surveyed reports passing NIST SP 800-22 and/or Diehard once appropriate post-processing, such as XOR combination, shrinking rules, or CRT-style reconstruction, is applied. Statistical adequacy is therefore largely achievable across map choices; the differentiator across studies is instead throughput and hardware cost. FPGA-targeted designs report the highest throughput and lowest resource use but the least adversarial evaluation, while software- or simulation-only designs report larger key spaces and more exhaustive test-suite coverage at the cost of speed. This is the same speed-versus-assurance trade-off quantified later in Table 8 and Figure 7.

Synthesis: Taken as a family, chaos-based PRNGs close much of the security gap left open by traditional generators without incurring the full computational cost of cryptographically secure designs. Their principal shared weakness is implementation rather than mathematics: finite-precision arithmetic, poorly chosen control parameters, and inadequate post-processing can silently degrade entropy or shorten the effective period. Nearly every strong result in Tables 4 and 5 therefore depends on an explicit precision- or post-processing-related design choice.

Hybrid and Residue-Number-System-Based PRNGs

Hybrid PRNGs combine linear generators such as LCGs with nonlinear chaotic maps. A general hybrid LCG, chaotic-map generator

Table 3 Comparative properties of selected chaotic maps used in PRNG design. "Chaotic range" denotes parameter values that produce well-characterized chaos

Map	Chaotic range	Advantages	Limitations	Applications
Logistic	$r \in (3.57, 4]$	Simple; fast; high sensitivity	Finite precision collapses chaos; insecure alone	Image encryption; keys; simulation
Tent	$\mu \in (1, 2]$	Uniform distribution; piecewise linear	Periodic if poorly tuned	Hardware RNGs; secure communications
Sine	$\mu \in (0, 1]$	High entropy; smooth nonlinearity	Limited dynamic range; scaling needed	Encryption; random-bit generation
Chebyshev	$k \geq 2$	Strong mixing; high unpredictability	Trigonometric overhead	Chaotic cryptosystems
PWLCM	$p \in (0, 0.5)$	Good uniformity; reduced periodicity	Sensitive to digital quantization	Image encryption; FPGA PRNGs
Ikeda	$\mu \approx 0.9$	High-dimensional chaos; strong nonlinearity	Complex and slower	Hybrid PRNGs; secure streams
Hénon	$a = 1.4$, $b = 0.3$	Bivariate chaos; high entropy	Precision loss; parameter sensitivity	Image encryption; masking
Lorenz	$\sigma = 10$, $\rho = 28$, $\beta = 8/3$	Continuous chaos; strong unpredictability	Numerical solvers; heavy computation	Physical modeling; communications
Tinkerbell	$a = 0.9$, $b = -0.6013$	Strong chaos; complex attractor	Parameter- and quantization-sensitive	Pixel permutation; encryption
Chirikov	$K > 0.9716$	Excellent mixing; ergodic behavior	High floating-point cost	Secure communications

Table 4 Comparative analysis of PRNGs based on chaotic systems (Part A). NIST = NIST SP 800-22; DH = Diehard; ENT = ENT test suite

Author (year)	Map(s)	Key findings and methodology	Strengths	Limitations
Al-Daraiseh et al. (2023)	Modified Tent (SBTM)	Circular state array and irrational seeds; 97.4% DH pass rate (88,830/91,200).	Flexible initialization; near-zero failure rate.	Large state arrays degrade performance.
De Micco et al. (2021)	Continuous chaotic (ROSEUR-MOD)	Modified Euler discretization; all NIST and DH tests passed with reduced precision; FPGA-efficient.	Resource-efficient; reduced-precision capable.	Output quality depends on time-step.
Irfan et al. (2020)	HC-MRLM + CoC	Control of Chaos for uniform binary distribution; all 15 NIST tests passed.	Strong unpredictability; large key space.	CoC region selection not fully justified.
Wang and Cheng (2019)	Logistic	Fifteen-digit transform and modular reduction; large dynamic key space.	Flexible key space; strong statistical properties.	Complex processing pipeline; susceptible to cryptanalysis.
Yu et al. (2019)	Four-wing memristive + Bernoulli	RK4 integration and XOR post-processing; 62.5 Mbps FPGA throughput; all NIST tests passed.	Dual-entropy architecture; high throughput.	Evaluated only on a single FPGA platform; no adversarial analysis.
Ismail et al. (2018)	Tent + Logistic hybrid	XOR combination outperformed individual maps; passed NIST SP 800-22; lightweight implementation.	High sensitivity; lightweight design.	Simulation-based validation only; no hardware implementation.

Table 5 Comparative analysis of PRNGs based on chaotic systems (Part B)

Author (year)	Map(s)	Key findings and methodology	Strengths	Limitations
Machicao and Bruno (2017)	Logistic (deep-zoom)	Extracts the k rightmost digits of chaotic orbits; achieves statistical quality comparable to MT19937.	Excellent statistical properties; passes multiple randomness test suites.	Computationally intensive; conjugacy regions must be excluded.
Jiang et al. (2017)	Diffusive memristor (TRNG)	Uses stochastic threshold switching in Ag:SiO_2 devices; generated 76 million bits and passed all 15 NIST tests.	Ultra-low power consumption; no post-processing required.	Sensitive to pulse parameters; device lifetime limited to approximately 10^7 cycles.
Avaroglu (2017)	Two Arnold cat maps	Employs sampler, comparator, and bit generator; successfully passes NIST tests with hardware-oriented implementation.	Low computational complexity; suitable for hardware implementation.	Sampling process decreases effective bit generation rate; lacks cryptanalysis.
Stoyanov and Kordov (2015)	Two Tinkerbell maps	Uses preprocessed real fractions; provides a key space of 2^{183} and passes NIST, Diehard, and ENT tests.	Large key space; resistant to brute-force attacks.	Lower generation speed than conventional non-chaotic PRNGs.
Stoyanov and Kordov (2014)	Two Chirikov maps + shrinking rule	Implements the Jabri shrinking generator; achieves near-ideal entropy and passes all NIST tests.	Simple architecture with excellent statistical performance.	Requires careful selection of parameters K_1 and K_2.

is

$$X_{n+1} = (aX_n + c) \bmod m, \tag{23}$$

$$x_{n+1} = f(x_n), \tag{24}$$

$$R_n = X_{n+1} \oplus \left(\lfloor x_{n+1} \times 10^k \rfloor \bmod m \right), \tag{25}$$

where $f(x) = rx(1-x)$ for the logistic map, $\oplus$ denotes XOR, k controls chaotic scaling precision, and R_n is the final hybrid output.

Residue-number-system-based PRNGs use modular arithmetic over coprime moduli combined through the Chinese Remainder Theorem (De Micco *et al.* 2021; Irfan *et al.* 2020). A number X_n is represented as

$$X_n \equiv (x_{n,1}, \ldots, x_{n,k}) \pmod{(m_1, \ldots, m_k)}, \tag{26}$$

with each residue evolving independently,

$$x_{n+1,i} = (a_i x_{n,i} + c_i) \bmod m_i, \qquad i = 1, \ldots, k, \tag{27}$$

and the combined output reconstructed through

$$X_{n+1} = \left(\sum_{i=1}^{k} x_{n+1,i} M_i M_i^{-1} \right) \bmod M, \tag{28}$$

where $M = \prod m_i$, $M_i = M/m_i$, and M_i^{-1} is the modular inverse of M_i modulo m_i.

Synthesis: Hybrid and residue-number-system designs occupy the middle ground between traditional, cryptographic, and chaos-based generators. The LCG stage in Equation (23) preserves the throughput of a linear generator, the chaotic stage in Equations (24)–(25) breaks the linear recoverability of a standalone LCG, and the formulation in Equations (26)–(28) removes carry propagation, making this family attractive for FPGA and parallel-hardware acceleration. This is reflected quantitatively in Table 8 and Figure 7, where hybrid and residue-number-system designs sit closer to the high-speed, high-security region than either purely traditional or purely chaotic designs alone.

Machine-Learning-Based PRNGs

Machine learning provides two important contributions to PRNG research. First, models serve as evaluators that learn hidden patterns and biases beyond what classical statistical tests reveal. Second, machine-learning architectures can become components of new PRNG designs (Boanca 2024; Kumar *et al.* 2023). Supervised learning uses labeled data to learn a mapping $f(x) \to y$; unsupervised learning finds structure in unlabeled data; semi-supervised learning combines a small labeled set with a larger unlabeled one; reinforcement learning trains an agent by trial and error; and deep learning uses multilayer networks to learn hierarchical representations.

Literature review: Boanca (2025) investigated artificial neural networks for enhancing LFSRs and the Geffe generator, proposing a Geffe-learning approach with 100% accuracy, though limited to known LFSR degrees. Egan (2024) examined high-speed secure random-number co-processors for privacy-preserving machine learning, identifying inadequacies of standard PRNGs for cryptographic security. Almardeny *et al.* (2023) proposed an upside-down reinforcement-learning system grounded in generalized information entropy that produces high-quality random sequences without training data. Crespo *et al.* (2024) assessed random-number quality using convolutional and long short-term memory networks, achieving approximately 79% classification accuracy, with wavelet features outperforming Fourier-based inputs. Truong *et al.* (2019) employed a recurrent convolutional network to analyze classical-noise impact on a quantum random-number generator, demonstrating robustness of post-processing against machine-learning attacks. Nagy and Suciu (2021) investigated neural architectures for random-number verification across quantum and linear congruential generators. Antunes and Hill (2024) compared Mersenne Twister, Philox, PCG, and Mrg32k3a across Python, NumPy, TensorFlow, and PyTorch using the Big Crush test, noting that PyTorch's default generator fails certain statistical tests. Dahiya *et al.* (2024) showed that PRNG tampering attacks, including bit-flipping, can degrade certified model robustness while evading NIST tests. Aljohani and Al-Barhamtoshy (2025) demonstrated that a 6–30–20–2 neural network detects implicit biases in cryptographically secure PRNGs more effectively than NIST SP 800-22. Fischer (2018) showed that long short-term memory testing detects hidden flaws missed by Diehard tests, while Wen and Yu (2019) proposed a physically unclonable-function-integrated PRNG achieving only 52.6% prediction accuracy, demonstrating resistance while maintaining high entropy.

On the residue-number-system and hardware side, Mehrabi *et al.* (2021) proposed an RNS-based GLV-ECC core resistant to machine- and deep-learning side-channel attacks. Mukhtar *et al.* (2022) conducted an extensive machine-learning side-channel analysis of RNS-ECC and reported 95–99% key-recovery accuracy. Valueva *et al.* (2020) implemented convolutional layers through residue arithmetic on FPGAs, with reported hardware-cost savings of 7.86–37.78 times. Peng *et al.* (2020) proposed DNNARA, a hybrid opto-electronic accelerator using residue arithmetic and wavelength-division multiplexing, achieving a 19.3-fold speedup over graphics processors.

Figure 5 summarizes the dual, and at times adversarial, roles that machine learning plays in this literature: evaluating existing generators for hidden bias, attacking generators to recover state, and directly generating new candidate PRNGs.

Synthesis: Two consistent findings emerge. First, machine-learning evaluators repeatedly find structure that classical NIST and Diehard tests miss (Fischer 2018; Aljohani and Al-Barhamtoshy 2025), suggesting that machine-learning-based testing should complement rather than replace standard suites. Second, machine-learning generation and residue-number-system hardware accelerators achieve strong statistical quality but at substantially higher computational and training cost than other families, which is the trade-off quantified next.

COMPARATIVE ANALYSIS

Methodology for Quantitative Comparison

The normalized comparisons in Tables 8 and 9, and Figures 6–7, are literature-derived syntheses rather than results from a single controlled experiment. For each of the fourteen criteria in Tables 6 and 7, the reported numerical values or ranges from the primary studies reviewed in Sections – were collected by family. Where a study reported a range, the midpoint was used as its representative value; where only a qualitative claim was available, such as "passes all NIST tests," it was scored at the upper end of the corresponding criterion for that family. Values were then min–max normalized to a common 0–10 scale within each criterion across all three families, and the per-family scores in Figure 6 are the resulting family-level averages. This approach follows the same aggregative logic used in prior PRNG comparison studies (Antunes and Hill 2024; Antunes 2025). It is intended to expose relative ordering and trade-offs across families rather than assert precision at the level of individual decimal values.

Traditional, Chaos-Based, and Machine-Learning-Based PRNGs

Traditional PRNGs such as LCGs are fastest, at approximately 10^6 numbers per second, but have limited security or statistical quality. Chaos-based PRNGs use nonlinear dynamics to provide moderate speed with good statistical quality, with reported NIST compliance around 97–98%. Machine-learning-based generators can approach 99% NIST compliance but are the slowest, at approximately 10^3 numbers per second. This heterogeneity drives research toward hybrid solutions combining advantages of all types. Tables 6 and 7 provide a detailed side-by-side comparison across fourteen criteria.

Discussion of Tables 6 and 7: The tables show a near-inverse relationship between randomness quality/security and computational efficiency/energy cost as one moves from traditional to chaos-based to machine-learning-based PRNGs. Traditional generators dominate every efficiency-oriented row but trail on quality-oriented rows such as entropy and attack resistance. Chaos-based generators sit between the two extremes on almost every criterion, making them attractive for resource-constrained but security-relevant deployments such as Internet-of-Things nodes. Machine-

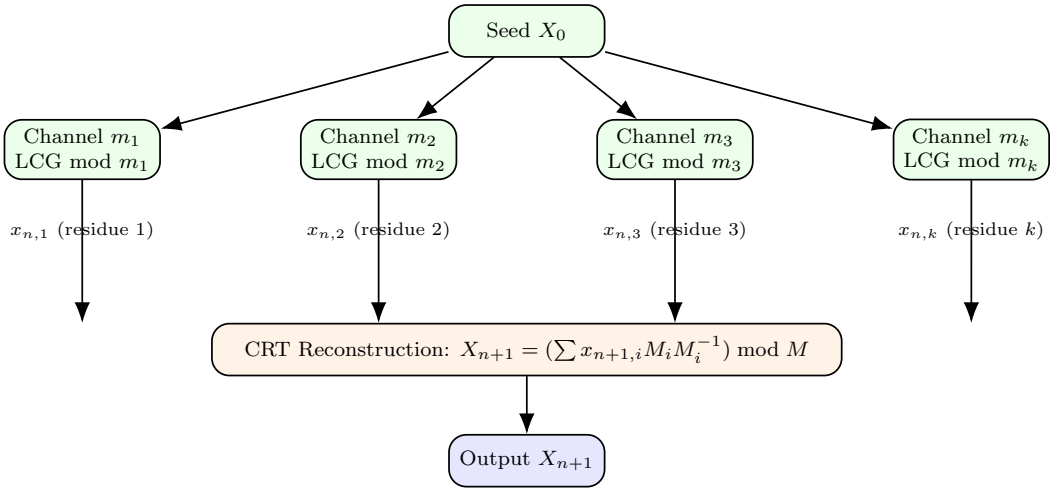


Figure 4 Residue-number-system-based PRNG architecture. The seed is decomposed into k independent residue channels; each evolves in parallel through a modular LCG recurrence; outputs are reconstructed into a full-period pseudo-random integer through the Chinese Remainder Theorem. This parallelism avoids carry propagation and enables hardware acceleration

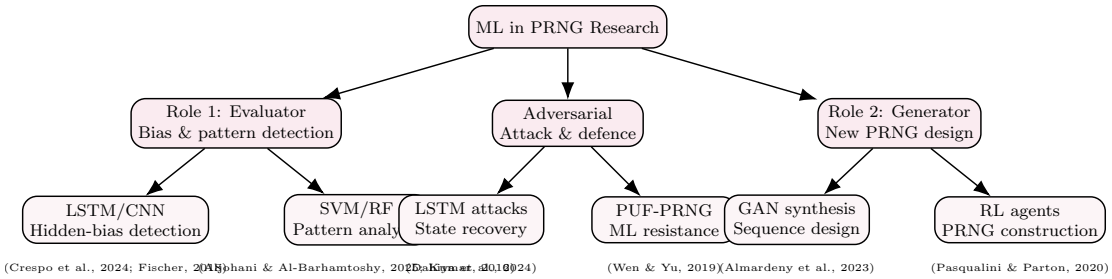


Figure 5 Dual and adversarial roles of machine learning in PRNG research. Machine learning contributes as an evaluator for bias and pattern detection, as an adversarial tool for state-recovery attacks and resistant design, and as a generator constructing new PRNG architectures through generative adversarial networks and reinforcement learning

Table 6 Comparison of Traditional, Chaos-Based, and Machine-Learning-Based PRNGs (Part A). ↑ = higher is better; ↓ = lower is better.

Criterion	Traditional PRNGs	Chaos-Based PRNGs	ML-Based PRNGs	Future ML Directions
Randomness quality ↑	Moderate; depends on seed quality	High; nonlinear sensitivity	Potentially very high with high-quality training data	Hybrid chaos–ML models; adaptive GAN-based entropy
Complexity ↓	Low; simple arithmetic operations	Moderate; nonlinear chaotic maps	High; computationally intensive training	Lightweight and edge-oriented ML architectures
Computational efficiency ↑	Very high; highly optimized implementations	Moderate; slower than conventional PRNGs	Lower due to training overhead	Optimized inference and edge deployment
Scalability ↑	Excellent in software and hardware	Moderate; synchronization required for coupled maps	High on GPU/cloud platforms; limited on embedded systems	Federated and distributed ML-based PRNGs
Security ↑	Limited unless cryptographically strengthened	High; unpredictable chaotic dynamics	Very high; neural parameters difficult to reconstruct	Adversarial learning and quantum-enhanced security
Adaptability ↑	Static after initialization	Moderate; parameter tuning possible	Highly adaptive through retraining	Self-learning PRNGs using reinforcement learning
Resource constraints ↓	Minimal CPU and memory usage	Moderate; floating-point arithmetic often required	High computational and memory requirements	Quantized neural models for IoT devices
Execution time ↓	Fastest	Fast to moderate	Slow due to model inference	Model pruning and hardware acceleration

learning-based generators top quality-oriented rows but perform worst on efficiency and hardware suitability; at present, they are therefore better suited to evaluation and offline generation than to real-time, high-throughput use.

Normalized Criterion-by-Criterion Comparison

Table 8 restates the comparison from a complementary angle, reporting representative numerical ranges for typical period, throughput, entropy, autocorrelation, NIST pass rate, security

Table 7 Comparison of traditional, chaos-based, and machine-learning-based PRNGs (Part B).

Criterion	Traditional PRNGs	Chaos-based PRNGs	ML-based PRNGs	Future ML Directions
Hybrid potential ↑	Common with secure generators and hashing	Very high; often hybridized	Emerging ML–chaos hybrids	ML–chaos–quantum frameworks
Entropy level ↑	Moderate (0.7–0.9 bits/bit)	High (≈1 bit/bit)	Very high (≈0.99 bits/bit)	Continuous-feedback entropy optimization
Period length ↑	Finite and known, up to $2^{19937}-1$	Long effective chaotic periods	Variable; architecture-dependent	Dynamic period control with recurrent ML
Autocorrelation ↓	Moderate linear dependency	Low chaotic correlation	Very low after training	Correlation-penalty loss functions
Uniformity ↑	Good with large moduli	Excellent with proper tuning	Near-perfect after calibration	Entropy-aware calibration networks
Hardware suitability ↑	Excellent for FPGA/ASIC	Moderate hardware cost	Limited by large model sizes	Efficient ML-PRNG accelerators
Energy consumption ↓	Low	Moderate	High during training and inference	Low-power inference; neuromorphic PRNGs
NIST/Diehard compliance ↑	Passes most tests; weak LCGs fail	Passes when parameters are optimal	Full compliance after training	Standardized ML-PRNG protocols
Attack resistance ↑	Low if seed is exposed	High; parameters difficult to predict	Very high; model parameters difficult to reverse	Adversarial robustness metrics

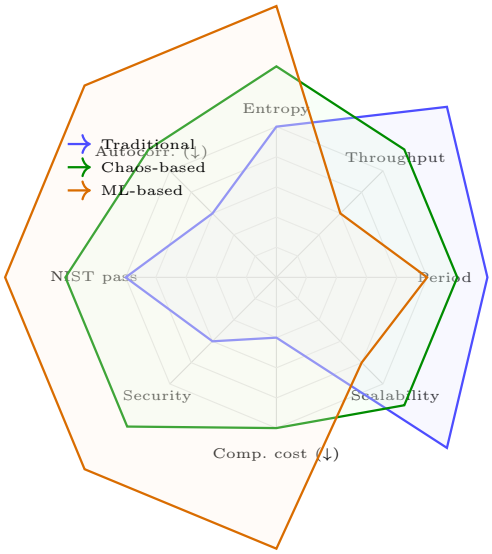


Figure 6 Normalized comparison of traditional, chaos-based, and machine-learning-based PRNGs across eight performance criteria on a 0–10 scale, derived using the aggregation method described in Section . Lower values are better for autocorrelation and computational cost

strength, computation cost, memory requirement, scalability, hardware suitability, determinism, and cryptographic applicability.

Discussion of Table 8: The absolute ranges reinforce the ordering seen in Tables 6–7 and Figure 6. Machine-learning-based PRNGs report the highest entropy and lowest autocorrelation but the lowest throughput and highest memory footprint. Traditional generators report the inverse. Chaos-based PRNGs again occupy the intermediate position on nearly every row, consistent with their use as a compromise choice for embedded and Internet-of-Things cryptography.

Security–Speed Trade-Off

The security–speed trade-off is a central design constraint in PRNG selection, and Table 9 summarizes it for four representative designs. LFSRs and Xorshift generators are extremely fast and hardware-friendly but linear, making them unsuitable for cryptographic use. MT19937 offers excellent statistical quality and a very long period but is not designed to resist state-recovery attacks. Hybrid PRNGs sacrifice some throughput for stronger nonlinearity, better entropy, and improved resistance to prediction. Figure 7 maps thirteen generators onto this two-dimensional landscape and confirms that cryptographically secure and hybrid/residue-number-system designs are the only families approaching the high-speed, high-security region simultaneously.

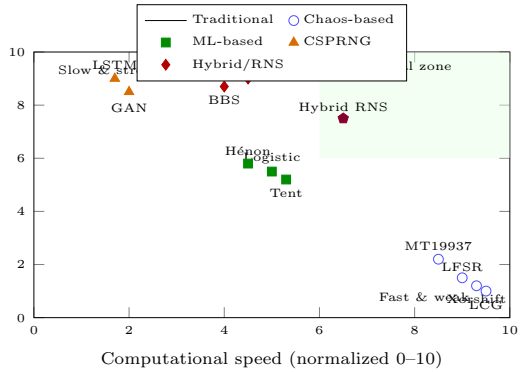


Figure 7 Security–speed landscape for 13 representative PRNGs, using the normalization methodology. Traditional generators cluster in the fast-but-weak region; chaos-based generators occupy the moderate zone; machine-learning-based PRNGs are strong but slow; cryptographically secure and hybrid/residue-number-system designs approach the desirable high-speed, high-security zone

DISCUSSION

The comparative findings indicate that PRNG performance is application-dependent. Traditional generators provide the best

Table 8 Normalized comparison of traditional, machine-learning-based, and chaos-based PRNGs

Metric	Traditional PRNGs (LCG, LFSR, MT19937)	ML-based PRNGs (DNN, GAN, RNN)	Chaos-based PRNGs (Logistic, Tent, Hybrid)
Typical period	10^6–10^{19}	10^{30}–10^{60}	10^{12}–10^{40}
Throughput (Mbit/s)	500–1500	50–200	200–1000
Entropy (bits/sample)	0.85–0.92	0.97–0.999	0.94–0.995
Autocorrelation	0.05–0.15	0.01–0.04	0.02–0.08
NIST pass rate (%)	85–90	97–99	93–98
Security strength (bits)	≤ 64 (weak)	≥ 128	96–128
Computation cost	Low	High	Medium
Memory requirement	16–256 bytes	5–50 MB	< 1 KB
Scalability	Excellent	Limited (GPU/TPU)	High
Hardware suitability	CPU, FPGA	GPU, TPU	CPU, FPGA, IoT
Determinism	Fully deterministic	Semi-deterministic	Deterministic
Cryptographic applicability	Weak–moderate	High	Moderate–high

Table 9 Security and speed trade-off for selected PRNGs

Type	Speed	Security	Trade-off Summary
LFSR	Very high	Very low	Fast but fully predictable through the Berlekamp–Massey algorithm; therefore, it is unsuitable for cryptographic applications.
MT19937	High	Low	Provides excellent statistical properties, but its internal state can be reconstructed from 624 consecutive outputs, making it unsuitable for cryptographic security.
Xorshift	Very high	Low	Extremely fast and computationally efficient due to linear operations; however, it lacks cryptographic security and is vulnerable to state recovery attacks.
Hybrid PRNG	Moderate–high	Moderate–high	Slightly slower than conventional linear generators, but offers significantly improved entropy, unpredictability, and robustness against statistical and differential attacks.

throughput and lowest implementation cost, making them suitable for simulation, graphics, sampling, and non-security-critical modeling; their weakness is predictability, since linear recurrence structures and recoverable internal states make them unsuitable for cryptographic key generation and secure communication. Chaos-based generators improve unpredictability through nonlinear dynamics, sensitivity to initial conditions, and low correlation, and are especially attractive for lightweight cryptography, image encryption, Internet-of-Things devices, and embedded security. Their main limitation is implementation fragility, since finite precision, poor parameter choices, and inadequate post-processing can reduce entropy or create short cycles.

Machine-learning-based approaches broaden the PRNG landscape by supporting both generation and evaluation: neural and reinforcement-learning methods detect hidden bias, model complex dependencies, and generate high-quality sequences, but they introduce significant computational cost, training dependence, reproducibility challenges, and hardware requirements. They are therefore best viewed as complementary tools rather than drop-in replacements. Hybrid and residue-number-system designs most directly address this three-way trade-off, and Table 8 and Figure 7 show them consistently closer to the jointly fast-and-secure region than either purely traditional or purely chaotic designs in isolation.

Future research on pseudo-random number generators should increasingly focus on integrative architectures that combine the computational efficiency of traditional generators, the nonlinear dynamics of chaotic systems, the parallel processing capabilities of residue-number systems, and the adaptive learning capacity of modern machine learning techniques. Such hybrid designs have the potential to simultaneously improve randomness quality, security, and computational performance. In particular, hybrid PRNG architectures integrating fast linear generators with chaotic or cryptographic post-processing can significantly enhance entropy and resistance against prediction attacks while maintaining high throughput.

Another promising direction involves the hardware acceleration of chaos-based and machine-learning-assisted PRNGs through FPGA, ASIC, GPU, TPU, and low-power embedded implementations, enabling their deployment in real-time and resource-constrained applications (Fang *et al*. 2014). Furthermore, precision-aware chaotic systems capable of preserving long periods, high entropy, and robust chaotic behavior under fixed-point arithmetic remain an important research challenge for practical hardware realization.

Future generators should also incorporate adaptive reseeding strategies using trusted entropy sources to maintain unpredictability even if part of the internal state becomes compromised. Finally, comprehensive evaluation methodologies should evolve beyond existing statistical frameworks by extending standards such as NIST SP 800-22, DIEHARD, and TestU01 to better assess the security and randomness characteristics of nonlinear, hybrid, and machine-learning-based PRNGs (Antunes 2025).

CONCLUSION

This review shows that PRNG design requires a careful balance among speed, randomness quality, predictability, memory cost, scalability, and cryptographic strength. Traditional generators remain valuable for high-speed, non-secure applications. Chaos-

based generators offer stronger nonlinear behavior and higher entropy but require careful parameter tuning and precision-aware implementation. Hybrid and residue-number-system methods provide a promising bridge between these families. Machine-learning-based approaches add value by detecting hidden patterns and supporting adaptive generation, although computational cost and reproducibility remain open challenges. No single PRNG family is optimal for every environment: simulation benefits from fast traditional generators; lightweight cryptography and Internet-of-Things applications from chaos-based or hybrid generators; and high-security systems from machine-learning-enhanced evaluation and adaptive reseeding. The most promising direction is a layered architecture combining speed, nonlinearity, entropy management, hardware efficiency, and rigorous testing.

Funding

This research received no specific grant from any funding agency, commercial entity, or not-for-profit organization. The study was conducted independently and fully supported by the authors.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research.

Availability of data and material

The data supporting the findings of this study are available within the manuscript. No external datasets were generated or analyzed. Additional information is available from the corresponding author upon reasonable request.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Al-Daraiseh, A., Y. Sanjalawe, S. Al-E'mari, S. Fraihat, M. Bany Taha, *et al.*, 2023 Cryptographic grade chaotic RNG based on tent-map. Journal of Sensor and Actuator Networks **12**: 73.

Aljohani, M. A. and H. Al-Barhamtoshy, 2025 A method for assessing the performance of cryptographic PRNGs utilizing ANNs. In *Proceedings of ICAISC 2025*, pp. 1–5.

Almaraz Luengo, E. and J. Román Villaizán, 2023 Cryptographically secured pseudo-random number generators. Mathematics **11**: 4812.

Almardeny, Y., A. Benavoli, N. Boujnah, and E. Naredo, 2023 A reinforcement learning system for generating instantaneous quality random sequences. IEEE Transactions on Artificial Intelligence **4**: 402–415.

Aniszewska, D., 2018 New discrete chaotic multiplicative maps based on the logistic map. International Journal of Bifurcation and Chaos **28**: 1850118.

Antunes, B. and D. R. C. Hill, 2024 Random numbers for machine learning. Journal of Data Science and Intelligent Systems .

Antunes, B. A., 2025 Statistical quality and reproducibility of PRNGs in ML. arXiv preprint arXiv:2507.03007.

Avaroglu, E., 2017 Pseudorandom number generator based on arnold cat map. Turkish Journal of Electrical Engineering and Computer Sciences **25**: 633–643.

Boanca, S., 2024 Exploring patterns and assessing the security of PRNGs with ML. In *Proceedings of ICAART 2024*, pp. 186–193.

Boanca, S., 2025 PRNGs, perfect learning and model visualization with neural networks. In *Proceedings of IoTBDS 2025*, pp. 117–128.

Crespo, J. L., J. González-Villa, J. Gutiérrez, and A. Valle, 2024 Assessing the quality of RNGs through neural networks. Machine Learning: Science and Technology **5**: 025072.

Dahiya, P., I. Shumailov, and R. Anderson, 2024 Machine learning needs better randomness standards. In *Proceedings of the 33rd USENIX Security Symposium*.

De Micco, L., M. Antonelli, and O. A. Rosso, 2021 From continuous-time chaotic systems to PRNGs. Entropy **23**: 671.

Egan, S., 2024 High-speed secure RNG co-processors for privacy-preserving ML. Unpublished manuscript.

Fang, X., Q. Wang, C. Guyeux, and J. M. Bahi, 2014 FPGA acceleration of a PRNG based on chaotic iterations. Journal of Information Security and Applications **19**: 78–87.

Fischer, S., 2018 Testing CSPRNGs with ANNs. In *Proceedings of SecITC 2018*.

Guisande, N., M. P. Di Nunzio, N. Martinez, O. A. Rosso, and F. Montani, 2023 Chaotic dynamics of the hénon map. Chaos **33**: 043111.

Irfan, M., A. Ali, M. A. Khan, M. Ehatisham-ul Haq, S. N. M. Shah, *et al.*, 2020 PRNG design using HC-MRLM. IEEE Access .

Ismail, S. M., M. N. Mohammed, M. A. Ameedeen, and H. A. S. Ahmed, 2018 A new trend of PRNG using multiple chaotic systems. Advanced Science Letters **24**: 7401–7406.

Jiang, H., D. Belkin, S. E. Savel'ev, S. Lin, Z. Wang, *et al.*, 2017 A novel true random number generator based on a stochastic diffusive memristor. Nature Communications **8**: 882.

Kumar, K. J., K. Jairam, and C. Ambedkar, 2023 An in-depth study of ML in artificial intelligence. Educational Administration: Theory and Practice .

Machicao, J. and O. M. Bruno, 2017 Improving pseudo-randomness properties of chaotic maps using deep-zoom. Chaos **27**: 053116.

Maheshwari, R., S. Gupta, V. Sharma, and V. Chauhan, 2014 In *Proceedings of IEEE IACC 2014*.

Martini, G. and F. G. Bruno, 2017 True random numbers generation from stationary stochastic processes. In *Proceedings of ICNF 2017*, pp. 1–4.

Mehrabi, M. A., N. Mukhtar, and A. Jolfaei, 2021 Power side-channel analysis of RNS GLV ECC. ACM Transactions on Internet Technology **21**: 1–20.

Mukhtar, N. *et al.*, 2022 ML-assisted side-channel attacks on RNS ECC using hybrid feature engineering.

Nagy, I. and A. Suciu, 2021 Randomness testing with neural networks. In *Proceedings of IEEE ICCP 2021*, pp. 431–436.

Noibate, S., 2023 Random number generators, challenges and limitations. SSRN Electronic Journal .

Oliveira, D. F. M., 2024 Mapping chaos: Bifurcation patterns and shrimp structures in the ikeda map. Chaos **34**: 123137.

Peng, J., Y. Alkabani, S. Sun, V. J. Sorger, and T. El-Ghazawi, 2020 DNNARA: A DNN accelerator using residue arithmetic and photonics. In *Proceedings of ICPP 2020*.

Saeed, H., W. I. El Sobky, T. O. Diab, and M. A. Elsisi, 2024 Chaotic maps in cryptography. In *Proceedings of ITC-Egypt 2024*, pp. 635–645.

Sathya, K., J. Premalatha, V. Rajasekar, P. Sathiyananthan, S. TharunPrasath, *et al.*, 2021 A cutting-edge approach to generate random bit sequences. AIP Advances **11**: 140006.

Sayed, W. S., A. G. Radwan, and H. A. H. Fahmy, 2016 Double-sided bifurcations in tent maps. In *Proceedings of ACTEA 2016*, pp. 207–210.

Silva, R. D. and S. D. Prado, 2023 Exploring transition from stability

to chaos through random matrices. Dynamics **3**: 777–792.
Stoyanov, B. and K. Kordov, 2014 A novel PRNG based on chirikov standard map. Mathematical Problems in Engineering **2014**: 986174.
Stoyanov, B. and K. Kordov, 2015 Novel secure PRNG based on two tinkerbell maps. Advanced Studies in Theoretical Physics **9**: 411–421.
Truong, N. D., J. Y. Haw, S. M. Assad, P. K. Lam, and O. Kavehei, 2019 Machine learning cryptanalysis of a quantum random number generator. IEEE Transactions on Information Forensics and Security **14**: 403–414.
Valueva, M. V., N. N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov, 2020 Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. Mathematics and Computers in Simulation **177**: 232–243.
Wang, L. and H. Cheng, 2019 Pseudo-random number generator based on logistic chaotic system. Entropy **21**: 960.
Wen, Y. and W. Yu, 2019 Machine learning-resistant PRNG. Electronics Letters **55**: 515–517.
Yu, F., Q. Wan, J. Jin, L. Li, B. He, *et al.*, 2019 Design and FPGA implementation of a PRNG based on a four-wing memristive hyperchaotic system. IEEE Access **7**: 181884–181898.

How to cite this article: Karim, M. A., Agbedemnab, P. A., and Alhassan, S. From Deterministic Chaos to Machine Learning: A Comparative Review of Pseudo-Random Number Generators for Modern Computing and Cryptography. *Chaos and Fractals*, 3(2), 125-136, 2026.