electronics

MDPI

Article

Tabular Foundation Models Versus Established Tabular Learners: A Regime-Stratified Critical Review

Huda Ali Alatawi

Department of Information Technology, Faculty of Computers and Information Technology, University of Tabuk, Tabuk 71491, Saudi Arabia; haladheedi@ut.edu.sa

Abstract

Tabular foundation models (TFMs) are pretrained once across many tables and applied to a new table by conditioning on its labelled rows, with no training on that table. They challenge gradient-boosted decision trees, the long-standing default for tabular prediction, but the published comparisons are hard to interpret, because studies differ in data regime, split design, baseline tuning, and metric. This review assesses the evidence regime by regime. From 76 peer-reviewed works selected under a documented protocol, it defines TFMs by reusable cross-table pretraining, organises them by pretraining data, context mechanism, scale, modality, and task interface, and compares them with tree ensembles, deep tabular networks, and tabular-to-image methods. Seven recurring claims are graded on peer review, evaluator independence, split match, baseline strength, and replication, with all scores published. One claim is strongly supported: on small-to-medium independent and identically distributed tables, TFMs match or exceed tuned tree ensembles without per-dataset tuning. Tuned tree-based and deep models remain as strong or stronger on temporal, feature-engineered data; above 10,000 rows the evidence is developer-reported except for one independent clinical benchmark; and no eligible study tests a frozen TFM under a grouped split or beyond its feature envelope. Tabular-to-image methods win only on small or very high-dimensional tables. Among the comparison-bearing studies, about one in ten use a deployment-matched split, none of them an independent evaluation of a TFM, and more than half of those evaluating a pretrained model report no contamination check. A reporting checklist, an evidence ladder, and seven research directions turn these findings into practice.

Keywords: tabular foundation models; in-context learning; prior-data fitted networks; gradient-boosted decision trees; tabular-to-image methods; trustworthy machine learning

Check for updates

Academic Editors: Jihoon Moon, Jehyeok Rew and Hyeon-Woo Kim

Received: 26 September 2026
Revised: 2 November 2026
Accepted: 2 November 2026
Published: 6 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Tabular data, heterogeneous values arranged in rows and columns, carry many high-value prediction problems in science and industry: clinical risk, credit scoring, materials discovery, environmental monitoring [1,2]. Deep learning has struggled with them. While neural networks came to dominate images and text, gradient-boosted decision trees (GBDTs) [3–5] stayed the default for tables, and repeated comparisons found tuned tree ensembles matching or beating neural networks [2,6,7]. Tabular machine learning therefore kept its per-dataset routine of feature engineering, model selection, and hyperparameter tuning.

Tabular foundation models (TFMs) change this routine by reusing one pretrained learner across table schemas. In the pure in-context family, the model receives the la-

belled rows of a new table as context and predicts the unlabelled rows in one forward pass (Figure 1). TabPFN made this concrete by pretraining on millions of synthetic datasets so that the network approximates Bayesian inference in context [8]. Its successor handles classification and regression on datasets of up to 10,000 rows and, in the evaluated classification setting, outperformed in 2.8 s an ensemble of baselines tuned for four hours [1]. Later models scale in-context learning to hundreds of thousands of rows and beyond [9–11], pretrain on real tables or on mixtures of priors [12,13], and add column semantics and free text [14–17].

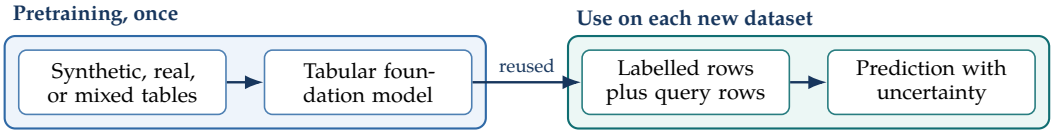


Figure 1. The tabular foundation model workflow: pretrained once across many tables, then reused by conditioning on the labelled rows of a new dataset.

The evidence for TFMs is strong but depends on the regime. Benchmarks place their advantage on small-to-medium independent and identically distributed (IID) tasks, tuned tree-based and deep models stay strong on temporal, feature-engineered data [7,18], and independent clinical benchmarking finds TFMs competitive but not consistently better than established methods [19]. Comparing such results is difficult. Studies vary in the data they use and how they split it, in how strongly and at what budget the baselines are tuned, and in context construction, preprocessing, and metrics. Several of these choices often go unreported, and results also depend on benchmark coverage, pretraining provenance, and contamination checks [17,20,21]. A headline gain may reflect a better learner or a more favourable evaluation, so the evidence has to be graded, not only organised. The review is built around one question: *where do tabular foundation models hold a supported advantage over established tabular learners, what evidence would establish it, and what does the current literature support?*

The review covers supervised classification and regression on single tables, where TFMs compete directly with tree ensembles, deep tabular networks, and tabular-to-image methods. Extensions to causal inference, forecasting, graphs, optimisation, and decision making are covered only as far as needed to fix the boundary of the evidence. The same search expressions retrieve them (Section 2.2), and they are graded as one claim (C5 in Section 5.5). The pure in-context line anchors the taxonomy, but adaptable, representation-transfer, and language-mediated models meet the same reusable-pretraining criterion and are treated as TFMs.

The review makes four contributions.

1. *A protocol-assembled evidence base.* A corpus of 76 works selected under stated questions, search expressions, inclusion criteria, and an extraction schema. It comes with an explicit membership rule for the 32 comparison-bearing studies that carry the graded conclusions, fixed regime thresholds, and an account of the content-relevant evidence set aside under the peer-review criterion (Section 2).
2. *A definition, a taxonomy, and a cross-paradigm comparison.* One criterion that separates TFMs from adjacent paradigms, four adaptation families, and five design axes that decide which comparison is fair for a given model. On these axes, a comparison of TFMs with tree ensembles, deep tabular networks, tabular-to-image methods, and hybrid tabular–visual networks (Sections 3, 4 and 6).
3. *Graded claims and testable explanations.* Seven recurring claims scored on a five-dimension rubric, with the dimension scores printed beside every grade and a sensitivity analysis, so that each grade can be recomputed and stress-tested. Candidate explanations for

the observed differences are stated as hypotheses with the experiment that would decide each (Sections 5 and 6).

4. *Two reusable instruments and a disclosure audit*. A TFM Reporting Checklist, audited against the 32 comparison-bearing studies of the corpus, and a TFM Evidence Ladder applied to representative models, with seven research directions (Sections 5, 7 and 9).

Section 2 sets out the review method, Section 3 the definition, background, and position among existing surveys, and Section 4 the taxonomy and model families. Section 5 synthesises and grades the benchmark evidence, Section 6 compares the paradigms, Section 7 covers trustworthiness and deployment, Section 8 sets out the application evidence, and Section 9 presents the cross-cutting discussion and research agenda.

2. Review Method

The evidence base was assembled under a written protocol that fixes the review questions, sources, search expressions, eligibility criteria, extraction schema, and grading rubric. Every field-level conclusion is tied to an evidence grade and to its supporting and countervailing sources.

2.1. Review Questions

Four questions follow from the central question of Section 1. Each fixed which fields were extracted from the reviewed works, and the conclusions answer them in turn (Section 10).

Q1 What separates a TFM from an adjacent paradigm, and how do TFMs differ among themselves? (Sections 3 and 4)

Q2 Under which data regimes do TFMs outperform tuned tree-based, deep, tabular-to-image, and AutoML baselines, and under which do they not? (Sections 5, 6 and 8)

Q3 Which recurring claims does the evidence support, and at what grade? (Section 5)

Q4 What evidence is required before a TFM is used in a high-stakes deployment? (Sections 7 and 9)

2.2. Sources, Search Expressions, and Selection

Records were identified from Google Scholar, arXiv, Semantic Scholar, DBLP, and the ACM, IEEE, Springer, and Elsevier indexes with five expressions (Table 1), adapted to each interface. The expressions target paradigm terms (S1), named models and benchmarks (S2), evaluation and trustworthiness qualifiers (S3), the baseline literature (S4), and tabular-to-image, spatial-encoding, and hybrid tabular–visual methods (S5). Searches cover records available up to 15 October 2026. The reference lists of the model, benchmark, and comparison papers, and the works citing them, were also searched, because application and trustworthiness studies are spread across domain venues and are poorly retrieved by paradigm terms. The protocol has two versions. The first fixed the review questions, expressions S1–S4, the eligibility criteria, and the grading rubric; the second added S5 and content criterion (iv) when the comparison was extended to tabular-to-image methods, before the searches were run. S5 was run on all eight sources under the same screening rules, so the analyses it supports, principally Section 6.1 and claim C7, can be traced to it. The protocol was not registered.

A record was included if it met at least one of four content criteria and the evidence criterion. The content criteria are: (i) it pretrains a model intended for reuse across heterogeneous tabular datasets; (ii) it benchmarks, adapts, analyses, or applies such a model; (iii) it supplies a tree-based, deep, or AutoML baseline against which TFM claims are tested; or (iv) it proposes or benchmarks a tabular-to-image or hybrid tabular–visual method for

supervised prediction on tables. The evidence criterion (v) requires a peer-reviewed version of record, a journal article or archival conference paper, which is also the version cited; throughout the review, eligible means satisfying this criterion. Records were excluded if they used tabular data only incidentally, proposed a per-dataset model with no cross-table pretraining and no comparator role, or had no English full text. A record superseded by a later version from the same authors gave way to that version. The publication status of every work was checked against publisher records at submission and again at revision.

Thirty-three records met a content criterion but existed only as preprints, technical reports, or non-archival workshop papers and were set aside under criterion (v). At revision, four of them had acquired a peer-reviewed version of record, one under a changed title, and these four are included and cited in their reviewed versions. The remaining 29 stay excluded and uncited (Table S3). They comprise fifteen vendor model releases or developer-attested methods with favourable findings, six studies that compare no methods, three with mixed findings, two independent favourable comparisons, two with unfavourable findings, and one regime-stratified benchmark with regime-dependent findings. Section 5.1 states which conclusions they bear on.

Table 1. Search expressions applied to the eight bibliographic sources.

Stream	Target	Expression
S1	Paradigm	(``tabular foundation model'' OR ``prior-data fitted network'' OR ``tabular in-context learning'' OR (``in-context learning'' AND tabular)) AND (classification OR regression OR prediction)
S2	Named models and benchmarks	TabPFN OR TabICL OR TabDPT OR Mitra OR CARTE OR TabSTAR OR ConTextTab OR TabuLa-8B OR MotherNet OR TabFlex OR LimiX OR TabArena OR BeyondArena OR TALENT OR TabReD OR OmniTabBench
S3	Evaluation and trustworthiness	(tabular AND (``foundation model'' OR ``in-context'')) AND (calibration OR ``uncertainty quantification'' OR ``proper scoring'' OR contamination OR memorisation OR fairness OR ``temporal split'' OR ``distribution shift'' OR interpretability OR ``fine-tuning'')
S4	Baseline literature, used only under criterion (iii)	(``gradient-boosted'' OR XGBoost OR LightGBM OR CatBoost OR AutoGluon OR ``deep tabular'') AND (benchmark OR comparison)
S5	Tabular-to-image and hybrid tabular–visual methods, used under criterion (iv)	(tabular OR table OR ``tidy data'' OR ``non-image'') AND (image OR images OR ``spatial encoding'') AND (``convolutional'' OR CNN OR ``vision transformer'' OR ``hybrid neural network'' OR ``deep neural network'' OR ``deep learning'' OR classification OR prediction)

Route labels (Table 2) are assigned from titles alone. A title takes the first expression, in the order S1, S2, S4, S5, whose paradigm phrase, named model or benchmark, baseline term, or table-and-image word pair it contains; matching ignores case, punctuation, and plural endings, and the task conjunct of S1 is not applied. S3 is a qualifier stream and is not used as a label.

Screening had three passes (Figure 2): deduplication across sources, title-and-abstract screening against the content criteria, and full-text assessment, at which point records found by citation searching were entered. Record counts at the identification and screening stages were not retained, so Figure 2 describes the structure of the selection and reports the counts that were retained; Table 2 gives the corpus by category and route. Because per-record retrieval logs were not kept, the route is a post hoc title classification, not a record of retrieval, and it cannot demonstrate the completeness of the search. It follows the mechanical rule stated beneath Table 1, so every label can be reproduced. A title with no expression term marks a work reached on full bibliographic records or by citation searching. The four known citation-searching entries are the paper that founds the prior-data fitted network lineage [22] and three tabular-to-image methods [23–25].

Reporting follows PRISMA 2020 [26] where the retained records allow. Risk of bias and reporting bias (items 11, 14, 18, and 21) are not assessed with a separate tool; the independence, split-match, and baseline-strength dimensions of the grading rubric and the account of the excluded records take their place. Effect measures and pooled synthesis (items 12, 13d, and 13f) do not apply, because reported quantities are transcribed, not pooled. Identification-stage counts (item 16a) and a registered protocol (item 24) are not available. The remaining items are reported in Sections 1–9 and Tables 1–4.

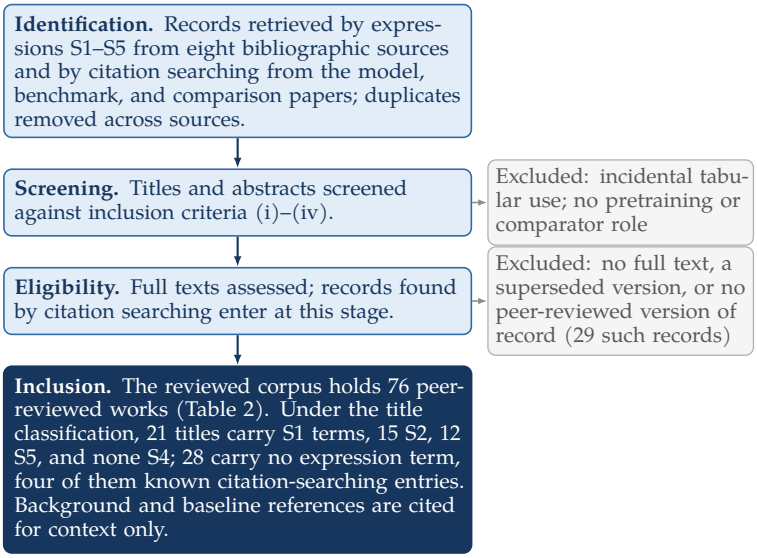


Figure 2. Structure of the selection procedure, with the counts retained at the inclusion stage.

Table 2. Composition of the reviewed corpus by category and by title-classified route.

Category or Route	Works
By category	
TFM model papers and the prior-data fitted network foundation	24
Task-interface extensions	12
Benchmarks and evaluation resources	6
Trustworthiness and mechanism studies	5
Adaptation, context, and serialisation methods	3
Tabular-to-image and hybrid tabular–visual methods	15
Adjacent surveys and position papers	6
Domain application studies	5
By title-classified route	
S1: paradigm terms	21
S2: named models and benchmarks	15
S4: baseline literature under criterion (iii)	0
S5: tabular-to-image, spatial-encoding, and hybrid terms	12
No expression term in the title	28
Reviewed corpus	76

Task-interface extensions cover survival, causal, time-series, stream, decision, and optimisation interfaces. Routes follow the rule beneath Table 1; the one baseline-literature work carries no S4 term in its title.

2.3. Data Extraction, Grading, and Synthesis

Every included work was read against the extraction schema of Table 3, which is weighted towards the variables along which reported results diverge: data regime, split design, baseline strength, tuning budget, and metric family [7,18,20]. Category, title-classified

route, and publication status are recorded for all 76 works and summarised in Table 2. For the comparison-bearing studies, the extraction adds the evidence role, the split design, whether a pretrained model is evaluated, and the contamination or leakage handling, which together fix the denominators of the audit in Section 5.4. The regimes, comparators, and budgets of the principal studies are given in Section 5.1. A work enters the matrix if it reports a quantitative evaluation and is cited as evidence in the comparison, claim, or application tables (Sections 5.1, 5.5, and 8); the two PFN theory papers [22,27] are cited there for a definition or a theoretical result and stay outside. Under this rule, the evidence matrix holds 32 studies (Table S1) and covers every work that carries a graded conclusion.

Table 3. Extraction schema for reviewed works.

Field Group	Extracted Items
Identity	Work type (model, benchmark, adaptation, application, trustworthiness, tabular-to-image, survey); venue; publication status; year.
Model	Pretraining data source; context mechanism or layout rule; adaptation family; supported rows, features, and classes; availability and licence.
Task	Task interface; output type; whether probabilistic output is produced and evaluated.
Evidence	Datasets and regime (IID, temporal, grouped, scale, dimensionality, text-richness); split design; baselines and their tuning budget; metrics; contamination or leakage checks.
Use-time	Context-selection rule; preprocessing; ordering or batching; adaptation; compute budget and hardware.
Trust	Calibration, subgroup, drift, privacy, or interpretability analysis; reproducibility artefacts.

Claims are graded with the fixed rubric of Table 4. A candidate claim qualifies when at least three works of the corpus assert or contest it and it bears on Q2–Q4; seven claims qualify. The supporting and the countervailing evidence of each claim are scored separately on five dimensions, each from 0 to 2: peer-review status, evaluator independence, split match, baseline strength, and independent replication. Split match rewards the design matched to the claim's regime, so a random split scores fully for a claim scoped to IID data and a deployment-matched split for a deployment claim. A source is independent for a claim when none of its authors, on the published author lists, is an author of a model whose performance the claim asserts. A replication is an evaluation of the same claim by a team outside the side's originating dependency groups. A developer source attesting to its own model does not replicate a claim about that model, and evaluations that share a benchmark ecosystem, meaning common datasets, code bases, or developer teams, count once. Three such groups recur: the TabPFN developer team with the arena benchmark and the time-series extension [1,20,28]; the team behind the two IIoT intrusion studies [29,30]; and the group behind the TINTO encoding, its library, the spatial-encoding benchmark, and two hybrid-network studies [31–35]. A source whose findings divide by regime enters each side with the finding that belongs to it, as the Gaussian-process comparison does for C3. No finding counts twice within a side.

The supporting total S fixes the band and the countervailing total K decides whether a claim is Contested, under the thresholds of Table 4. Equal weights are the neutral default of an additive appraisal index, the Strong band demands near-complete credit with no absent dimension, and Section 5.5 reports how every grade behaves under alternative weights and thresholds. A Contested label means substantial evidence on both sides and is reported with both side bands, so Strong support against Strong counter-evidence stays distinguishable from Moderate on both sides. Countervailing evidence outside a claim's stated scope fixes that scope and does not lower the grade. When it narrows a claim, the narrowing is written into the claim and the scope-fixing sources stay listed in the claim table of Section 5.5, so a grade cannot rise by silent narrowing.

Screening, extraction, and scoring were carried out by the author under this protocol, so the grades are author-assigned assessments under the published rubric rather than consensus judgements. The coding rules are stated with the rubric and the dimension scores are printed beside every grade (Sections 5.5 and 8; per-source scores in Table S2), every count derives from the extracted records, and the sensitivity analysis bounds the effect of the scoring choices. Because every cited source meets criterion (v), the peer-review dimension separates sides only by their number of sources. Extracted records were synthesised thematically in two tiers: Sections 4–8 analyse each evidence domain, and Section 9 reads the domains against one another. Version-sensitive statements are tied to named releases and dated publication records.

Table 4. Grading rubric applied to the supporting and the countervailing evidence of each claim.

Dimension	**Score 0**	**Score 1**	**Score 2**
A. Peer review	Preprints, reports, or workshop papers only	One peer-reviewed source	Two or more peer-reviewed sources
B. Independence	Developer-affiliated sources only	Independent sources support part of the claim	Independent sources support the whole claim
C. Split match	Split unstated	Split stated but not matched to the claim's regime	Split matched to the claim's stated regime
D. Baseline strength	Default or unstated tuning	Tuned baselines without a stated budget	Tuned baselines at a stated budget
E. Replication	None	One independent team replicates	Two or more teams replicate
Band	**Preliminary**	**Moderate**	**Strong**
Side total S (of 10)	$S \leq 4$	$5 \leq S \leq 7$, or $S \geq 8$ with a zero dimension	$S \geq 8$ with no zero dimension
Contested	Assigned when the supporting total S and the countervailing total K both reach 5, and reported with both side bands. Counterevidence outside a claim's stated scope fixes that scope and does not enter K.		

Peer-reviewed means a journal article or archival conference proceedings; replication is defined in Section 2.3.

3. Definition, Background, and Position

3.1. Definition and Adaptation Families

A *tabular foundation model* is a model pretrained once, at scale, across many heterogeneous real or synthetic tabular tasks, so that it provides a learning capability reusable on unseen table schemas without redesign. The defining criterion is reusable cross-table pretraining. It separates TFMs from models fitted separately for each dataset and carries the foundation-model idea, one pretrained model adapted to many tasks, to tables.

Within this class, models differ in how they specialise to a new dataset. Four adaptation families result: (i) *pure in-context* models that predict by conditioning on context rows with no downstream parameter updates (the PFN/ICL line, the reference case throughout); (ii) *adaptable* models that add fine-tuning, adapters, retrieval, or learned contexts to a pretrained learner; (iii) *representation-transfer* models that pretrain reusable table, column, or entry representations and fit a light downstream head; and (iv) *language-mediated* models that process serialised tables through a pretrained language model. Families (ii)–(iv) meet the same criterion, which is why CARTE, TabSTAR, and TabuLa-8B appear among the model families of Section 4. Models that do not meet it are adjacent paradigms: per-dataset GBDTs and deep networks, per-dataset representation learners, tabular-to-image pipelines, generation-only models, and prompt-only use of a general language model. Tabular-to-image methods and hybrid tabular–visual networks fit their image layout and vision network per dataset, but they compete for the same prediction problems and enter the comparison of Section 6 under criterion (iv), which places every paradigm on common axes.

Three borderline cases fix the boundary. "Without redesign" means without a new architecture per schema, so adapting parameters, heads, or prompts on the new table does not disqualify a model. CARTE is a representation-transfer TFM, because its pretrained graph representation transfers to any schema and only a light head is fitted. TabLLM is a language-mediated TFM, because the pretrained language model supplies the reusable capability while serialisation and few-shot adaptation specialise it. MotherNet is a pure in-context TFM whose hypernetwork emits the dataset-specific predictor in the forward pass, so the per-dataset object is an output of the pretrained model, not a retrained one.

3.2. Tabular Data and the Persistence of Tree Ensembles

A tabular problem supplies rows of heterogeneous features (numerical, categorical, sometimes free text) with labels and asks for predictions on unseen rows. Unlike pixels or tokens, tabular features have no natural order or locality, signals are often sparse, irregular, and heavy-tailed, datasets are often small, and little structure transfers across datasets [2,6]. GBDTs handle these properties cheaply, which explains their persistence. Neural approaches, from attentive architectures such as TabNet [36] to feature tokenisers, retrieval-augmented designs, and efficient deep ensembles such as FT-Transformer, TabR, and TabM [37–39], narrowed the gap on some benchmarks, and well-regularised simple networks are competitive [40]. Broad comparisons nevertheless found that they did not clearly displace tuned trees [6,7,41], and strong tuned baselines remain hard to beat [42]. This review compares two responses. Tabular-to-image methods keep the per-dataset workflow and give the data an artificial spatial structure so that vision networks can apply (Section 6.1); TFMs remove the per-dataset workflow and amortise learning itself into one pretrained network.

3.3. Prior-Data Fitted Networks and In-Context Learning

In-context learning (ICL), first shown at scale by large language models [43], lets a pretrained model adapt to a task from examples in its input, with no parameter updates. For tables, the labelled training split is the prompt: the model conditions on those rows and predicts the query rows in one forward pass [1,9]. In common implementations, the fit step stores or prepares the context and the predictive forward pass runs at inference time [1].

The substrate for tabular ICL is the *prior-data fitted network* (PFN) [22]. One specifies a prior over datasets, samples many synthetic datasets from it, and trains a transformer to predict held-out labels from the rest of each dataset, minimising the expected log loss over datasets and over randomised context–query splits. The trained network then approximates the posterior predictive distribution of a query label given its features and the context: the likelihood under each data-generating hypothesis in the prior, averaged with weights given by how well that hypothesis explains the context. The integral is never computed: the network emulates it [22,27], and statistical analysis has begun to characterise when the emulation is sound [27]. Because the output is a predictive distribution, PFNs provide uncertainty estimates, whose reliability Section 5.3 examines.

A PFN is only as good as its prior. TabPFN samples synthetic datasets whose feature–label relations come from structural causal models with randomised structure and parameters [1]. Synthetic pretraining avoids direct exposure to private or licensed tables, lowers benchmark-overlap risk, and lets designers shape the inductive bias by editing the prior [1,44]. Its limit is coverage: a synthetic prior may miss real data mechanisms, missingness patterns, rare categories, and domain semantics. Real, continued, and mixed pretraining respond to this limit (Section 4.3).

Two complementary readings of the in-context computation exist. The *Bayesian* reading is the original one: a PFN approximates the posterior predictive distribution, so its predic-

tions depend on the prior it was trained on [22,27]. The *mechanistic* reading inspects the trained network: analyses of TabPFN v2 identify randomised feature tokens as important for handling heterogeneous schemas and probe when the model simplifies the downstream distribution [45]. The readings describe one computation at different levels of abstraction, and both are consistent with the documented dependence of outcomes on which rows enter the context [46]. Their shared consequence is that behaviour depends on the prior, on the similarity structure of the context, and on architectural choices, so context construction is part of the scientific claim (Section 4.2).

3.4. Position Among Existing Reviews

Table 5 places the review among existing surveys and benchmark contributions. Deep tabular learning surveys treat TFMs as one recent architecture family, not as a distinct workflow [2,47]. Representation-learning surveys are closest in spirit, but their object is the learned representation, not the in-context workflow, its use-time choices, and the evidence needed to trust its outputs [48]. Generation surveys ask how to generate data, not how a pretrained model should solve a new prediction task [44]. Benchmark papers supply the strongest primary evidence, and this review depends on them [18,20,49]. Each is scoped to its own protocol, however, which is why their disagreements call for a synthesis, and the spatial-encoding benchmark of Liu et al. [33] plays the same role for tabular-to-image methods. Three gaps remain. No review takes the pretrained in-context learner as its unit of analysis. No synthesis states, claim by claim, what the combined benchmark evidence supports. And use-time design, the context a model is given, is treated as an implementation detail although it changes accuracy, calibration, and fairness [46]. This review addresses all three.

Table 5. Coverage of this review and of existing surveys and benchmark contributions.

Work Family	Unit of Analysis	TFM Focus	Regime Strata	Paradigm Comparison	Graded Claims	Context Protocol	Trust and Deployment
Deep tabular learning surveys [2,47]	Per-dataset architectures	○	○	◐	○	○	◐
Representation-learning surveys [48]	Learned representations	◐	○	○	○	○	○
Structured-data foundation models [50]	Tables, time series, graphs	◐	○	○	○	○	○
Tabular generation surveys [44]	Synthetic data	○	○	○	○	○	◐
Position paper [51]	Research agenda	●	○	○	○	○	○
Benchmark contributions [18,20,33,49]	One protocol each	◐	●	◐	○	○	○
This review	Pretrained in-context learner	●	●	●	●	●	●

● Central to the work; ◐ Partly covered; ○ Not covered.

4. Design Taxonomy and Model Families

4.1. Five Design Axes

We organise TFMs along five design axes (Table 6): pretraining data, architecture and context mechanism, scale regime, modality, and task interface. The axes determine how a model is trained, what it can use at inference, which datasets it handles, and which evaluation claims are admissible.

Pretraining data fixes the prior the model brings to a new table. Synthetic priors give control over the learning problem and reduce dependence on private or web-collected tables, at the cost of coverage. Real-data and representation-transfer approaches such as TabDPT and CARTE expose the model to real tables or cross-schema structure [12,16];

mixed strategies such as Mitra combine the two [13]. The choice affects accuracy, privacy, licensing, and the credibility of benchmark claims.

Table 6. Design taxonomy: five axes, each a spectrum read left to right, with the consequence each choice has for evaluation.

Design Axis	Choice Spectrum	Representative Models	Consequence for Evaluation
Pretraining data	Synthetic prior → mixed prior → real-table corpora → web-scale text corpora	TabPFN; Mitra; TabDPT; TabuLa-8B	Whether contamination can be audited or only asserted; privacy and licensing exposure
Architecture and context	Full attention over rows and features → column-then-row encoding → linear attention → hypernetwork → retrieval	TabPFN v2; TabICL; TabFlex; MotherNet; TabDPT	Whether the whole training table is usable, hence whether a context-selection rule becomes part of the model
Scale regime	≤1k rows → ∼10k → ∼500k → ∼1M rows, each with its own feature ceiling	TabPFN v1/v2; TabICL; TabICLv2; TabFlex	The comparator: a small-data claim competes with defaults, a large-table claim with tuned GBDTs and AutoML at matched budget
Modality	Numeric and categorical arrays → column names and units → entity strings and free text	TabPFN line; ConTextTab; CARTE; TARTE; TabSTAR	Whether provenance, serialisation sensitivity, and leakage checks are required before a result is interpretable
Task interface	Classification and regression → density and survival → causal and time series → decisions and optimisation	TabPFN v2; CausalFM; Do-PFN; GIT-BO	The admissible metric: accuracy and AUC suffice only for the first group; the rest require proper scores or task-native validation

Architecture fixes how context becomes prediction. TabPFN v2 uses attention across features and examples; TabICL uses a column-then-row encoder so that larger contexts are practical; TabFlex uses linear attention; MotherNet uses a hypernetwork that emits a dataset-specific predictor [1,9,11,52]. These choices decide whether the model can use the full training table, whether it must select a context subset, and whether inference cost grows smoothly or sharply with rows and features.

Scale is a regime, not a single row count. A reported limit depends on feature count, class count, hardware, context construction, ensembling, batching, and test-time compute. The comparator also changes with scale: a small-data TFM is valued for removing tuning, while a large-table TFM must be compared with tuned GBDTs, deep models, and AutoML under comparable budgets [9,10].

Modality fixes what information can enter the model. Early PFN-style systems treated tables as numeric and categorical arrays. Newer systems add column names, entity strings, and free text: CARTE represents entries and columns through a graph of context-aware entries, TabSTAR and ConTextTab use semantic information, and TabuLa-8B serialises tables for a language model [14–17]. Semantics widen the range of tasks and change the risk profile: provenance, leakage, serialisation sensitivity, and privacy need stronger checks.

Task interface links the design to the evidence a claim needs. Classification and regression remain the core use, and the same amortised-inference idea has been adapted to survival analysis, causal inference, time series, reinforcement learning, Bayesian optimisation, and graph node prediction [28,53–61] (Table 7). Each interface needs its own validation: proper scoring rules and calibration for densities, censoring-aware evaluation for survival, explicit assumptions about treatment assignment and confounding for causal effects, and policy-level validation for decisions. A study that evaluates only classification provides no evidence for regression, density modelling, causal estimation, or survival prediction, and a method proposed for uncertainty-aware prediction must report probabilistic metrics, not only accuracy or AUC.

The axes interact: a semantics-aware model benefits from real-data pretraining but faces stronger provenance questions, a large-context model needs a specialised architecture

or batching strategy, and a causal or density interface needs different metrics. Before asking whether one model is best, the reader should therefore identify the pretraining source, context mechanism, scale regime, modality, and task interface being evaluated.

Table 7. Task interfaces of TFMs and the output each requires.

Interface	Required Output	Representative Work
Classification	Class probabilities	TabPFN, TabICL, TabDPT [1,9,12]
Regression	Mean and uncertainty	TabPFN v2, TabDPT [1,12]
Density estimation	Conditional density; log likelihood; CRPS	PFN posterior predictive output [22,27]
Survival analysis	Survival curve or time-dependent risk	Survival-aware adaptation [53]
Causal inference	Treatment effects	CausalFM, CausalPFN, Do-PFN [54–56]
Time series and streams	Forecasts or stream predictions	ForecastPFN, TabPFN extensions [28,57,62,63]
Decisions and optimisation	Policies, values, or next candidates	PFNs4BO, TabPFN-based RL, GIT-BO [58–60]
Graph node prediction	Node-level predictions	GraphPFN [61]

4.2. Context Construction and Use-Time Choices

For a TFM, downstream use begins before the forward pass. A table has to be turned into a context: labelled examples, query rows, feature encodings, optional schema text, batching choices, and sometimes adaptation steps. These choices change accuracy, calibration, fairness, leakage risk, and runtime even when no gradient training takes place, so training-free use is not choice-free use.

Six decisions carry most of this weight: context selection (size, sampling rule, class and group balance, seeds, and no use of test labels [46]); ordering and batching (row order, chunking, order ensembling, context limit [1]); preprocessing (missingness, categorical encoding, scaling, feature filtering, leakage checks); schema semantics (whether names or text were used, serialisation format, backbone, provenance [14–16,64]); adaptation (frozen or adapted, validation budget, objective, calibration before and after [65,66]); and compute allocation (hardware, inference and tuning time, ensemble size, test-time compute, baseline search budget). They form one disclosure group of the reporting checklist (Section 5.4), and two studies of the same TFM evaluate different workflows if any of the six differ.

Context selection is the central decision. If the model can use the whole training table, the context is the training set; if only a subset fits, the selection rule becomes part of the model. Nearest-neighbour, class-balanced, random, and uncertainty-selected contexts expose the model to different local patterns and group proportions, which matters for subgroup reliability and fairness as well as accuracy. The rule should therefore be fixed before test labels are seen and reported in reproducible detail. Preprocessing stays part of the claim for the same reason, above all when TFMs are compared with GBDTs and deep networks, which benefit from different defaults. Adaptation changes what a number measures: a frozen TFM measures the pretrained learner and its context mechanism, whereas a fine-tuned, adapted, retrieved, ensembled, or distilled TFM measures a broader workflow. The two should not be mixed without stating validation and compute budgets.

4.3. Model Families

The TFM literature reads best as a set of responses to a few bottlenecks: context size, prior coverage, schema meaning, and interface breadth. Figure 3 places the main milestones on a timeline coloured by this reading, and Table 8 maps representative models to the design axes.

The TabPFN family is the reference point because it made tabular ICL concrete. TabPFN v1 showed that a transformer pretrained on synthetic, SCM-generated datasets could solve small classification problems in context [8]. TabPFN v2 extended the idea to regression with alternating column and row attention and reported strong results up to roughly 10,000 rows at much lower per-task compute than tuned ensembles [1]. Later releases of the lineage exist only as vendor technical reports, so they fall outside the corpus under criterion (v), and the lineage is assessed on its peer-reviewed members.

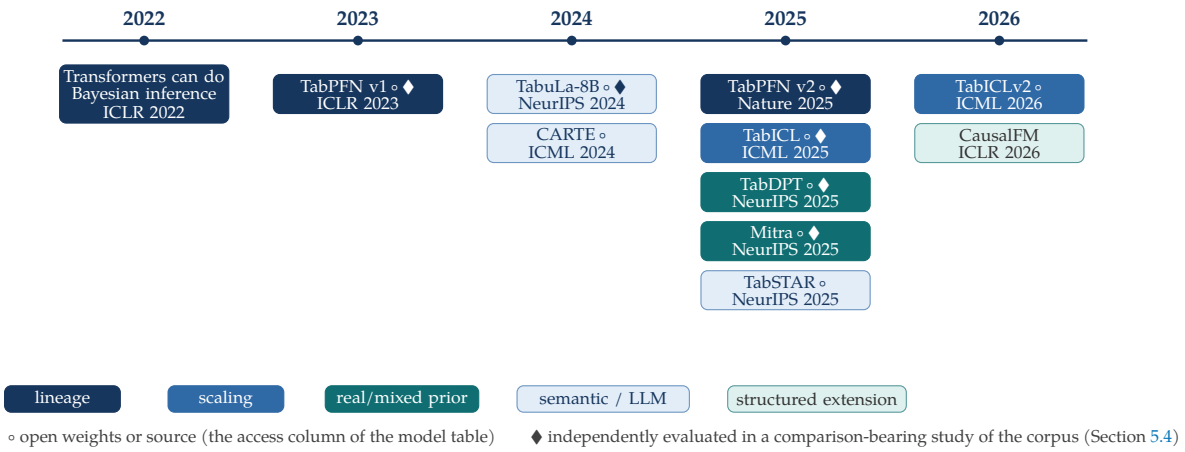


Figure 3. Milestones in the tabular foundation model lineage, coloured by design theme, with openness and independent-evaluation markers; dates and venues follow the cited publication records.

Table 8. Representative TFMs mapped to the design taxonomy.

Model	Pretraining Data	Context Mechanism	Tasks	Reported Scale	Access
PFN reference lineage and architectural variants					
TabPFN v1 [8]	Synthetic SCM prior	Transformer over rows	Classification	≤1k rows, 100 features, 10 classes	Open weights
TabPFN v2 [1]	Synthetic	Attention across features and rows	Classification, regression	≈10k rows, 500 features	Open weights
MotherNet [52]	Synthetic	Hypernetwork emits a predictor	Classification	Small tables	Open source
Scalable in-context models					
TabICL [9]	Synthetic	Column-then-row encoder	Classification	≈500k rows	Open weights
TabICLv2 [10]	Synthetic	Improved column-then-row encoder	Classification, regression	≈1M rows under 50 GB	Open weights
TabFlex [11]	Synthetic	Linear attention	Classification	Millions of rows	NR
Real and mixed pretraining					
TabDPT [12]	Real tables	Retrieval-based ICL	Classification, regression	Retrieval-scaled	Open weights
Mitra [13]	Mixed synthetic priors	TabPFN-style ICL	Classification, regression	Small to medium	Open weights
Semantic, representation, and language-mediated transfer					
CARTE [16]	Real cross-table corpora	Graph of entries; fine-tuning	Classification, regression	Small to medium	Open source
TabSTAR [14]	Real tables with text	Text encoder with target tokens	Text-rich classification	Medium to large	Open source
ConTextTab [15]	Real tables with semantics	Semantics-aware ICL	Classification, regression	Small to medium	Open source
TabuLa-8B [17]	Web tables	Serialised rows in Llama-3-8B	Zero- and few-shot classification	Context-limited	Open weights

NR: Not reported by the source; SCM: Structural causal model.

Scaling work starts from a practical constraint: full-context ICL becomes costly as the training table grows, so larger TFMs change either the architecture or the way context is processed. TabICL operates on up to roughly 500,000 samples, although it was pretrained

on synthetic datasets of at most 60,000, and it matches TabPFN v2 across the 200 TALENT classification datasets while running up to ten times faster [9]. TabICLv2 reports million-scale operation under 50 GB of GPU memory [10], and TabFlex targets million-row settings with linear attention [11]. TabSwift targets inference cost instead of context size and reports accuracy competitive with TabPFN v2 and TabICL from a row-wise attention-only backbone with an adaptive early exit [67]. Adaptations rescue small-context models on larger data: TuneTables compresses large datasets into learned contexts [65], retrieval-based localisation fine-tunes on nearest-neighbour contexts [66], and aligned encoders extend TabPFN to large, high-dimensional tables [68]. Larger context capacity is informative only when the evaluation also reports cost, batching rules, calibration, and the regime in which the scaling claim was obtained.

The choice of pretraining data is the main split among the families. Synthetic-prior models control the pretraining distribution and avoid overlap with public benchmark tables [1]. TabDPT pretrains on real tables with self-supervised learning and retrieval [12], Mitra mixes synthetic priors [13], and XTab explored cross-table pretraining before the current wave [69]. Real and mixed pretraining cover realistic schemas better but need provenance, deduplication, licensing, and leakage controls, so pretraining data is an evaluation variable, not only a development detail.

Semantics-aware and representation-transfer models address the meaning that column names, units, free-text fields, and entity values carry and numeric arrays discard. TransTab pretrains transferable row encoders that carry over across tables within a domain [70]. TabSTAR includes native text fields [14], ConTextTab uses semantic information inside an in-context learner [15], CARTE represents tables as graphs of context-aware entries so that transfer across schemas needs no explicit column matching [16]. TARTE reuses column and entry semantics across datasets [64]. Language-mediated models form the fourth family. TabLLM serialises rows into text prompts [71], TP-BERTa pretrains a language model for tabular prediction with relative-magnitude tokenisation of numeric values [72], MediTab uses language-model assistance to consolidate heterogeneous medical tables into reusable domain predictors [73], and TabuLa-8B fine-tunes a Llama-3-8B model on a large table corpus for zero- and few-shot transfer [17]. These models matter most when schema language and text fields are central to the task. They differ from PFN-style TFMs in context limits, input biases, privacy concerns, and contamination risk, because public web tables can overlap with evaluation data [21].

Architectural variants and structured-task extensions complete the picture. MotherNet produces a dataset-specific predictor in one forward pass [52], and HyperFast meta-trains a hypernetwork of the same family that emits a ready classifier for an unseen table [74]. Equi-TabPFN makes the architecture equivariant to target permutations [75], and Drift-Resilient TabPFN replaces the exchangeable-row prior with one whose data-generating process drifts over time [76]. Beyond prediction, PFN and ICL machinery has been adapted to tabular generation [44], causal inference [54–56], time series and evolving streams [28,57,62,63], survival analysis [53], reinforcement learning [58], Bayesian optimisation [59,60], and graph node prediction [61]. The maturity of evidence differs sharply across these branches (Section 5.5).

5. Evaluation Evidence

5.1. Regimes of Advantage and Parity

Table 9 consolidates the principal evaluation studies and codes, for each regime a study evaluates, which paradigm the study's own primary aggregate favours (mean rank, Elo, or dataset win count), or Mixed when it reports no consistent leader. Cells transcribe the study's result and never re-analyse it. Table 10 gives the quantitative anchors the

studies report, including effect sizes and win–tie–loss counts. TabPFN v2 reports strong small-data classification and regression results with large speed advantages over heavily tuned ensembles [1], and TabArena finds deep and foundation-model approaches highly competitive on smaller datasets, especially with larger inference or ensembling budgets [20]. These results make TFMs serious competitors, not universal replacements.

Regime terms carry fixed thresholds. Small-to-medium means at most 10,000 training rows and large means more than 10,000, following the TabPFN v2 envelope and the TALENT size split. High-dimensional means more than 500 features, the same envelope's feature ceiling. The temporal, grouped, and text-rich strata follow the split or schema structure of the data. Under these thresholds, three kinds of evidence have to be kept apart. Operation at scale is demonstrated in developer evaluations under random splits: TabICL leads TabPFN v2 and CatBoost on the 53 TALENT datasets above 10,000 rows and operates to roughly 500,000 rows [9]. TabICLv2 and TabFlex report operation to a million rows [10,11]. Independent evaluation above 10,000 rows is confined to the clinical benchmark, whose cohorts reach 139,528 patients, although it reports one pooled result across cohort sizes and evaluates the v2.5 release of the reference model [19]. Deployment-matched and grouped evaluation is where eligible coverage ends. No eligible study evaluates any TFM under a grouped split. The one temporal-shift evaluation in the lineage is the developer study of Drift-Resilient TabPFN. It pretrains on priors that drift over time and reports higher accuracy and better calibration than XGBoost, CatBoost, the frozen reference model, and the Wild-Time baselines [76]. Beyond 500 features, the eligible evaluations adapt the model with trained encoders instead of using it frozen [68].

Established models keep their ground elsewhere. TabReD shows that industry-style temporal splits and feature-engineered data reorder the rankings of deep networks and GBDTs, although it evaluates per-dataset models only [18]. The broadest meta-feature analysis finds that whether neural or tree-based models win tracks dataset properties such as size and feature count, with no family dominating [7], and text-rich, high-cardinality schemas are addressable only by semantics-aware or language-mediated models [14,15,17]. Independent clinical benchmarking finds TFMs competitive on several tasks while established methods stay comparable or stronger in important settings [19]. The conclusion is regime-based. TFMs suit small-to-medium IID problems with low tuning budgets. Tuned GBDTs and neural baselines remain the necessary comparators for temporal, grouped, and beyond-envelope data, where eligible TFM evidence is developer-reported or absent, and text-rich schemas need a semantics-aware or language-mediated model before the comparison means anything. The empty cells of Table 9 sit exactly where TFMs are least established.

The 29 excluded preprint-only records bear on these conclusions in stated ways. The broadest regime-stratified benchmark evaluates eleven models on 142 datasets spanning IID, temporal, and grouped tasks, sample sizes, dimensionalities, and text features. It finds TFMs ahead on small-to-medium IID data and tree-based and deep models ahead in the non-IID, large, and high-dimensional regimes. Admitting it would fill the temporal, grouped, and large-scale cells of Table 9 in the direction stated above and supply the regime-stratified evidence the Evidence Ladder requires. The vendor report of the reference lineage claims operation to 50,000 rows and 2000 features with developer-attested results and a distillation engine that bears on Direction D7. The other vendor releases make similar claims, so admitting them would extend the favourable record at scale without adding independence. An independent credit-risk comparison reports untuned TFMs ahead of established methods, with gains growing as datasets shrink, which would strengthen C1 inside its regime. The remaining records concern prior design, memorisation, fine-tuning, and inference cost, and bear on C4, C6, D4, and D7. No graded conclusion reverses

under their admission; the main change would be from absent to stratified coverage of the non-IID regimes.

Table 9. Principal empirical comparisons of TFMs with established baselines, with the paradigm each favours in every regime it evaluates.

Source	Scope, Comparators, and Split	IID	Temporal	Grouped	Large	High-Dim.	Text	Uncert.	Principal Result and Limit
TabPFN v2 [1]	Up to 10k rows; tuned ensembles and AutoML at a four-hour budget; random splits	■							Beat the tuned ensemble in 2.8 s; larger and shifted data out of scope
TabICL [9]	200 TALENT classification datasets, 53 above 10,000 rows; TabPFN v2 and CatBoost among the comparators; random splits	■			■				Ahead of TabPFN v2 and CatBoost on the 53 datasets above 10,000 rows; developer-reported
TabArena [20]	Living leaderboard; 51 small-to-medium tasks; TFMs, deep models, tuned GBDT ensembles; random splits	■							TFM and deep methods lead at larger inference budgets; mostly IID
Clinical benchmark [19]	Twelve binary tasks, 788 to 139,528 patients, results pooled over sizes; TabPFN (v2.5 release) against twelve methods	■			■				Most AUROC gaps within ±0.02; best baseline beaten in 16.7% of tasks; 5.5 × runtime
Drift-Resilient TabPFN [76]	18 synthetic and real datasets under temporal shift; XGBoost, CatBoost, frozen TabPFN, Wild-Time baselines		■						Accuracy 0.688 to 0.744 and ROC AUC 0.786 to 0.832 over the frozen model; developer-reported
Uncertainty benchmark [77]	112 TALENT datasets; four TFMs against GBDTs; random splits	■						■	Highest AUC; weaker conditional coverage than GBDTs when stratified by conformal set size
Regression with GPs [78]	Regression problems of varying size at 4 to 40 features; TabPFN against Gaussian processes	■			■			■	Competitive with sufficient data; Gaussian processes often ahead when data are scarce

■: TFM favourable; ■: Mixed; ■: Tree-based or deep favourable; ■: Not evaluated. IID: Small-to-medium IID tables; Large: More than 10,000 training rows; High-dim.: More than 500 features; Text: Text-rich features; Uncert.: Quality of predictive distributions. TabPFN v2, TabICL, and Drift-Resilient TabPFN are developer evaluations, and TabArena shares authors with the TabPFN developer team; the clinical, uncertainty, and Gaussian-process benchmarks are independent of the model developers. The clinical benchmark reports one pooled result across its twelve cohorts, so its IID and Large cells carry the same finding.

Table 10. Quantitative anchors reported by the principal comparisons, consolidated from the studies' own figures.

Source	Regime or Target	Reported Quantity
TabPFN v2 [1]	Small IID classification	Beat an ensemble of baselines tuned for four hours in 2.8 s
TabICL [9]	200 TALENT datasets	On par with TabPFN v2 at up to 10 × lower runtime; ahead of TabPFN v2 and CatBoost on the 53 datasets above 10,000 samples
Clinical benchmark [19]	Twelve clinical tasks	AUROC gaps mostly within ±0.02; best baseline beaten in 2 of 12 tasks (16.7%); 5.5 × median runtime
Uncertainty benchmark [77]	112 TALENT datasets	Highest AUC with weaker set-size-stratified conditional coverage than GBDTs
Regression with GPs [78]	Regression at 4 to 40 features	Competitive with sufficient data; Gaussian processes often ahead on accuracy and uncertainty when data are scarce
Spatial-encoding benchmark [33]	24 datasets, tuned GBDTs	Classical and neural pipelines win 12 datasets each; layout fitting up to 10^4 s at 1776 features

Each entry restates the cited study's own reported figures. GPs: Gaussian processes.

5.2. Sources of Disagreement Between Benchmarks

Different benchmarks ask different questions. Random-split OpenML-style suites [79] and the historical neural-versus-tree comparisons measure general predictive performance but over-represent small IID tasks, the regime in which current TFMs match or exceed

tuned trees [7]. TALENT is a unified evaluation toolbox for deep tabular learning, not a deployment stress test [49]. TabReD emphasises industry-style temporal splits and feature-engineered data [18], and TabArena maintains a living leaderboard of TFMs, deep models, and tuned tree ensembles on curated small-to-medium tasks [20]. Each claim therefore needs the benchmark that can test it: small-data speed, temporal generalisation, and text-rich schemas call for different evidence. The last case is critical for language-mediated models, because web-scale pretraining can overlap with public evaluation tables and large language models memorise common public datasets [17,21,80]. The comparison-bearing evidence also leans towards classification, on which TALENT and the arena suites centre; eligible regression evidence comes from the model papers' own evaluations and the comparison with Gaussian processes [1,12,78].

5.3. Probability, Calibration, and Density Quality

Point accuracy does not suffice for TFMs. PFN-style models approximate posterior predictive inference and output predictive distributions [22,27], whose reliability has to be tested, and the eligible evidence consists of two comparisons that bound the question from opposite sides. Against default-setting Gaussian processes on regression problems of varying size at 4 to 40 features, TabPFN is competitive on the larger and more complex problems in that range given sufficient data, while Gaussian processes often give better accuracy and uncertainty estimates when data are scarce or the kernel matches the underlying function [78]. Across the 112 TALENT datasets, four TFMs achieve the highest AUC but weaker conditional coverage than gradient-boosted trees under conformal prediction, with coverage stratified by the size of the conformal prediction set, not by dataset size [77,81]. No eligible benchmark yet scores TFM predictive densities across many datasets under proper scoring rules, so density quality remains an open measurement task (Direction D2). Discrimination and reliability are therefore separate axes: accuracy, AUC, and RMSE suit point prediction, whereas probability claims need log loss, Brier score, calibration error [82], coverage, CRPS, or other proper scoring rules [83].

5.4. Minimum Evidence for a Reliable Comparison

A reliable comparison states the dataset regime, baselines, tuning budget, context construction, preprocessing, metrics, split design, contamination checks, and model version, because these decide whether two results measure the same claim. A frozen TFM against a lightly tuned GBDT answers a different question from an ensembled or adapted TFM against a high-budget AutoML pipeline. Random splits on temporal data overstate performance, and context rows chosen with information from test labels invalidate an in-context comparison. Figure 4 consolidates these requirements, with the artefact-identification and reproducibility details of Section 7, into the *TFM Reporting Checklist*. Every item is there because at least one reviewed study shows that it changes a reported result.

Two items of the comparison-protocol group can be audited directly from the evidence matrix, whose split and pretrained-model codes fix each study's membership in the two denominators (Table 11). Of the 29 studies for which a data split applies, 3 evaluate under a deployment-matched split (temporal or horizon-based): a benchmark that evaluates no TFM [18], a forecasting extension by the model developers [28], and a drift-adapted variant evaluated by its developers [76]. No independent study therefore evaluates any TFM under a deployment-matched split. None of the application studies, where deployment splits matter most, use one. Of the 24 studies that evaluate a pretrained model, 4 report an explicit contamination or leakage check or a provenance disclosure, 7 rely on the structural argument that a synthetic prior cannot overlap with evaluation data, and 13 report nothing. A study is coded as making the synthetic-prior argument only when its text invokes synthetic

pretraining as the reason overlap cannot occur, including a benchmark that evaluates only synthetic-prior models. A study that neither checks nor argues is coded as reporting nothing, whatever could be inferred about its models. Curation against known within-benchmark leakage, as the arena benchmark reports, is not an overlap check against a pretraining corpus and is coded as reporting nothing [20].

Table 11. Two checklist items applied to the 32 studies of the evidence matrix.

Item	Bench-marks	Models	Extensions	Applica-tions	All
Deployment-matched split (studies with a data split)					
Temporal, grouped, external, or hori-zon split	2 of 11	0 of 10	1 of 4	0 of 4	3 of 29
Contamination or leakage handling (studies of a pretrained model)					
Explicit check or provenance	0 of 8	4 of 6	0 of 5	0 of 5	4 of 24
Synthetic-prior argument only	4 of 8	1 of 6	2 of 5	0 of 5	7 of 24
Not reported	4 of 8	1 of 6	3 of 5	5 of 5	13 of 24
Studies in the group	8	6	5	5	24

Groups follow the evidence role coded in the matrix (benchmark, model, structured-task extension, application), not the work category of Table 2. Studies without a data split and per-dataset methods without a pretrained model are excluded from the respective denominators.

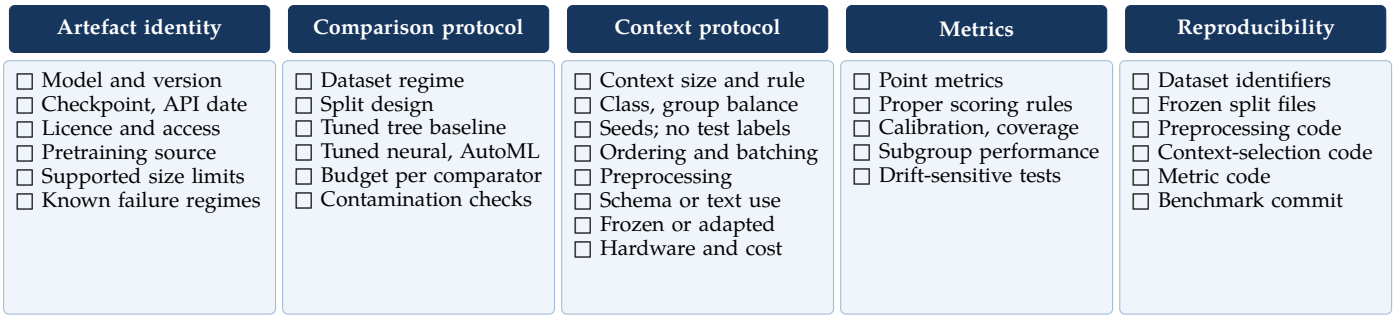


Figure 4. The TFM Reporting Checklist, grouped by the part of an evaluation each item documents.

5.5. Seven Recurring Claims, Graded

The evidence above distils into seven claims that recur in the literature. Table 12 grades each with the rubric of Section 2.3 and prints the dimension scores behind every grade. The strongest claims hold within specific regimes and interfaces and weaken when generalised. The small-data claim (C1) holds inside its regime and gives no support to a universal replacement claim. Its independent support comes from the two benchmarks author-disjoint from the model developers [19,77]; the arena benchmark also supports it but shares authors with the TabPFN developer team. The clinical benchmark reports its twelve cohorts only as a range, 788 to 139,528 patients, with cohorts above 100,000 sampled down to 100,000, and pools the results without stratifying by size, so the number of tasks inside the small-to-medium regime cannot be read from it. It therefore enters C1 as independent evidence of parity with a split-match score of 1, stated but not matched to the regime; the side's split-match score of 2 comes from the developer benchmarks, so the C1 totals are unchanged.

One claim is Contested, meaning that evidence at the Moderate band or above exists on both sides. C2 is graded at the scope a deployment relies on: that a frozen TFM needs no per-dataset tuning for accuracy, calibration, or subgroup outcomes. On accuracy, the claim is supported, and the fairness study supports it as well, because its context interventions leave accuracy nearly constant [46]. The same study shows that those interventions move group-fairness outcomes, so for outcomes other than accuracy the choice of context is a per-dataset decision that has to be made and reported. Scoped to accuracy alone, C2

would be Moderate with no countervailing source. The three adaptation studies in the corpus [65,66,68], each motivated by the row or feature limits of the frozen model, show that per-dataset adaptation raises accuracy over frozen use; they mark where the tuning-free regime ends (Section 5.1 and Direction D5) rather than contest the claim inside it, and are cited here as scope, not scored. For predictive distributions (C3), eligible support is thin and regime-bound while the independent counter-evidence on conditional coverage is strong, so the claim is Preliminary. Prior design (C4) is Moderate for the sufficiency of synthetic priors. The mixed-prior evidence is itself a mixture of synthetic priors [13], so the countervailing side rests on one developer study of real-data pretraining [12]. Among structured-task interfaces (C5), only classification and regression are mature, and any generalisation claim by a language-mediated model (C6) first needs contamination-controlled evaluation. The tabular-to-image advantage (C7) is Preliminary, because its five supporting method papers are developer-attested and none replicates another's claim. In the one benchmark that tests the family under one protocol, it is confined to small samples and the highest-dimensional tables studied.

The grades are stable under the scoring choices (Table 13). Six variants were recomputed: the Strong threshold at 7 or 9, the Contested threshold at 4 or 6, the independence dimension double-weighted, and no zero-dimension cap. Every grade is unchanged except under a Contested threshold of 4, where C3 becomes Contested because its counter-evidence is Strong, and C4 because its counter-evidence sits one point below the default threshold. No supporting band rises under any variant.

Table 12. Seven recurring claims in the TFM literature with their supporting and countervailing evidence, rubric scores, and grades.

Claim	Supporting Evidence	Countervailing Evidence or Scope Limit	Support Total S A/B/C/D/E	Counter Total K A/B/C/D/E	Grade
C1. TFMs match or exceed tuned GBDTs on small-to-medium IID tables	Strong results and large speed gains [1,20]; independent benchmarks report parity or better, pooled over dataset sizes [19,77]	Temporal, feature-engineered, and text-rich data set the scope [14,18]	10 2/2/2/2/2	0 in scope	Strong
C2. TFM use needs no per-dataset tuning for accuracy, calibration, or subgroup outcomes	One forward pass matches tuned baselines on accuracy [1,9]	Context selection moves subgroup-fairness outcomes at near-constant accuracy [46]	6 2/0/2/2/0	6 1/2/2/1/0	Contested (Moderate vs. Moderate)
C3. TFM predictive distributions are reliable off the shelf	Competitive distributions given sufficient data [78]	Weaker conditional coverage across 112 datasets [77]; Gaussian processes often ahead on scarce data [78]	4 1/1/2/0/0	8 2/2/2/1/1	Preliminary (Preliminary vs. Strong)
C4. Synthetic priors are sufficient	Synthetic-only models competitive without benchmark overlap [1,9]; a mixture of synthetic priors improves accuracy further [13]	Real-table pretraining improves accuracy [12]	6 2/0/2/2/0	4 1/0/2/1/0	Moderate (Moderate vs. Preliminary)
C5. TFMs extend to general structured-data inference	Causal, time-series, and decision extensions [28,54,58]	No interface yet carries an eligible validation independent of its own developers	4 2/0/2/0/0	—	Preliminary
C6. Language-model table predictors generalise broadly	One developer evaluation using schema text [17]	Memorisation and serialisation effects [21,80]	3 1/0/1/1/0	8 2/2/2/1/1	Preliminary (Preliminary vs. Strong)
C7. Tables rendered as images let vision networks beat tuned GBDTs	Gains on omics, drug-response, and benchmark tasks [23,84–87]	A 12–12 tie against tuned GBDTs, with wins confined to the data-scarce and highest-dimensional datasets [33]; converter-dependent [88]	3 2/0/1/0/0	8 2/1/2/2/1	Preliminary (Preliminary vs. Strong)

A–E: Rubric dimensions of Table 4; —: Counter-evidence limits scope only; (a vs. b): Support and counter bands. E counts replications by teams outside a side's originating dependency groups: two independent benchmarks for C1; one further independent team for the counter-evidence of C3, C6, and C7; none for C2, C4, C5, and C7's support, where every source attests its own model, prior, interface, or encoding.

Table 13. Grade stability of the seven claims under alternative scoring rules.

Scoring Rule	C1	C2	C3	C4	C5	C6	C7
Rubric of Table 4	Strong	Contested	Preliminary	Moderate	Preliminary	Preliminary	Preliminary
Strong threshold at 7 or at 9	✓	✓	✓	✓	✓	✓	✓
Contested threshold at 4	✓	✓	Contested	Contested	✓	✓	✓
Contested threshold at 6	✓	✓	✓	✓	✓	✓	✓
Independence double-weighted	✓	✓	✓	✓	✓	✓	✓
No zero-dimension cap	✓	✓	✓	✓	✓	✓	✓

A check mark means the grade is unchanged from the rubric grading in the first row.

6. Cross-Paradigm Comparison

Section 5 establishes where each family leads. This section compares the paradigms on common axes and states candidate explanations for the differences, keeping the findings of the cited work apart from the interpretations offered here. Table 14 sets seven paradigms against the dimensions the evaluation literature treats as decisive: the locus of learning, the regime in which the paradigm leads, behaviour under non-IID structure, and interpretability. Additive and depth-limited models occupy the interpretability corner and are the operative comparator in regulated settings. Tabular-to-image and hybrid methods compete for the same prediction problems and offer a useful contrast, because they manufacture the locality that tabular data lack, the structure that PFN-style priors supply synthetically.

Table 14. Tabular learning paradigms compared on the dimensions that determine where each leads.

Paradigm	Locus of Learning	Regime of Advantage	Non-IID Behaviour	Interpretability
Adjacent paradigms fitted per dataset				
Additive and depth-limited models [89]	Per dataset	Very small data; explanation as the deliverable	Visible degradation	Intrinsic and global
Tree ensembles [3,90]	Per dataset	Irregular targets; large, engineered tables [6,18]	Refit per split	Exact attribution [91]
Deep tabular networks [37,39,42]	Per dataset	Large tables with local structure	Refit per split	Post hoc; examples in retrieval designs [38]
Tabular-to-image and hybrid networks [33,84]	Per dataset, plus a fitted layout	Small samples or the highest dimensionalities tested [33]	Layout refit per split	Saliency; branch scores [35]
Tabular foundation models pretrained across tables				
Representation transfer [16,64,69]	Pretrained; head per dataset	Semantic columns and entities	Head refit per split	Post hoc
Language-mediated models [17,71]	Pretrained language model	Text-rich schemas; few-shot use [21]	Sensitive to serialisation [71]	Token level, after memorisation checks
Pure in-context TFMs [1,9,12]	Pretrained once; table as context	Small-to-medium IID, low tuning budget [1,20]	Exchangeable-row prior	Post hoc; mechanism probes [45]

6.1. Tabular-to-Image and Hybrid Tabular–Visual Methods

This family keeps the per-dataset workflow and changes the data instead of the learner: each row is rendered as an image so that convolutional networks or vision transformers can be applied. *Parametric layout methods* compute a feature-to-pixel layout from the training table so that similar features land on neighbouring pixels. DeepInsight, REFINED, IGTD, and TINTO differ in projection and refinement, and the TINTOlib library implements them behind one interface [23,31,32,84,85]; FC-Viz places correlated and interacting features side by side to obtain smaller images and faster networks at similar accuracy [92]. *Non-parametric methods* map values to the image by a fixed rule, from printed feature values to pixel areas proportional to feature importance [24,25]. *Integrated methods* learn the transformation jointly with the classifier [86,87]. *Hybrid tabular–visual networks* add an MLP branch on the raw features and fuse it with the image branch. The published examples are indoor-localisation networks reading MIMO channel tables, later extended to CNN–KAN hybrids

with branch-level attribution [34,35], and transfer learning across tabular-to-image tasks has also been tested [93]. Fusing real images with tables [94,95] is a different problem, in which the image is an input modality rather than an encoding of the table, and lies outside this review.

The method papers report gains on specific, mostly high-dimensional biological and chemical tasks against baselines each chose. The family's coverage is concentrated. The TINTO encoding and library, the spatial-encoding benchmark, and the two hybrid-network studies share a research group on their author lists [31–35], and the replication dimension counts them once. The benchmark of Liu et al. [33] is the first to evaluate the family under one protocol against tuned GBDTs; its authors developed the TINTO encoding and library it includes, so it is developer-affiliated for part of what it evaluates. It covers nine encodings, four vision or hybrid architectures, and 24 regression and classification datasets from the literature and the OpenML CC18 and CTR23 suites, with Optuna tuning for every model.

Its findings follow the regime structure of Section 5. Classical and neural pipelines each win 12 of the 24 datasets. Neural pipelines win in the data-scarce regime ($N \lesssim 5$k) and on the highest-dimensional datasets, from the 617-feature ISOLET data to the 1600-feature MIMO data, and GBDTs win on medium-to-large tables with low-to-moderate feature counts. The transformation has a larger and statistically significant effect on accuracy than the vision architecture. REFINED is the most robust encoding, leading a statistically indistinguishable group (IGTD, distance matrix, combination), and hybrid CNN + MLP and ViT + MLP models reduce predictive variance, most visibly on small datasets, while ViT, CNN, and hybrid variants are not statistically separable overall. Parametric encodings are costly at high dimensionality (REFINED takes in the order of 10^4 s to fit on the 1776-feature Bioresponse data), and non-parametric ones scale poorly because image size grows with feature count. Categorical features enter through one-hot encoding, which disrupts spatial coherence.

Two consequences follow for the comparison with TFMs. First, both families claim the small-sample regime, where a manufactured spatial prior or a learned prior supplies structure the data alone cannot. No eligible study compares them directly: the spatial-encoding benchmark includes no TFM, and the TFM benchmarks include no image pipeline; Direction D1 states the experiment. Second, they differ in what is reused. A tabular-to-image pipeline fits its layout, network, and hyperparameters per dataset, so it inherits the tuning cost of deep tabular networks plus the transformation cost and transfers nothing to the next table. A TFM pays its cost once, at pretraining. The practical trade is a per-dataset design search on one side against a fixed prior with a context protocol on the other.

6.2. Candidate Explanations for the Observed Differences

Four explanations account for most of the pattern in Table 14. The first is established by controlled experiments. The other three are hypotheses, each stated with what would confirm or refute it.

H1 (established): target geometry favours axis-aligned partitioning. Controlled comparison established three properties of typical tabular problems that penalise neural architectures: target functions are often irregular while neural networks favour smooth solutions; uninformative features degrade multilayer perceptrons more than trees; and tabular data are not rotationally invariant, so rotation-invariant architectures discard structure that axis-aligned splits keep [6]. This is a property of the data, not of the optimiser, and it explains why tuned trees stayed hard to displace across a decade of neural proposals [7,41,42]. Tabular-to-image methods are consistent with it, since they give the data a locality the original features lack.

H2 (hypothesis): the pretrained prior substitutes for data that are not there. A PFN approximates the posterior predictive distribution under a prior over datasets [22,

27]. If most of the predictive content on a small table comes from that prior, the advantage should be largest where per-dataset estimation is least reliable and should need no tuning, because the model averages over hypotheses instead of selecting one. The evidence fits: the advantage concentrates at small-to-medium scale and arrives without per-dataset tuning [1,20]. Explicit priors set a boundary, because default-setting Gaussian processes often beat TabPFN on accuracy and uncertainty in the same data-scarce regime [78]; the hypothesis therefore concerns a learned prior against no prior, not against a well-matched explicit one. The alternative is that the gain comes from the ensembling and preprocessing in the inference pipeline. Ablating the ensemble would separate the two, as would narrowing or widening the prior at fixed architecture while measuring the gap to tuned trees.

H3 (hypothesis): the prior's marginal contribution declines as the table grows. If H2 holds, the advantage should shrink as n grows and the likelihood comes to dominate the prior. The evidence supports a conditional form. The advantage concentrates at small scale and the winner tracks dataset properties including size [7], yet architectures built for larger contexts win at larger n, with TabICL ahead of both TabPFN v2 and CatBoost above 10,000 samples [9]. A competing explanation is context truncation: models that cannot hold the full table subsample it, and TabPFN v2 truncates at that scale, so the cited result cannot separate the two readings. Varying n at fixed context capacity, and comparing models whose capacity exceeds the table, would.

H4 (hypothesis): exchangeability is the binding assumption under non-IID structure. Pretraining fits the network over randomised context–query splits, which assumes exchangeable rows. A temporal table violates this because the admissible split is ordered, and a grouped table because it is by group. On this reading, the non-IID deficit is a mismatch between the pretraining objective and the deployment split, which should persist as models scale and close only when the prior or the context rule changes. Time-based validation that leaves simple MLPs and GBDTs strongest on industry-grade data is consistent with this [18]. The competing explanation is that shift hurts every model and TFMs merely lack a re-fitting step, which fine-tuning on the deployment distribution would supply. Direction D5 states the separating experiment: pretrain with an order-aware or group-aware prior and test, at fixed capacity and against a fine-tuned baseline, whether the temporal deficit closes. Drift-Resilient TabPFN runs the order-aware half of that experiment and reports a favourable developer result under temporal shift (Section 5.1) [76]. The group-aware half and an independent replication remain open.

A fifth difference is a channel, not a mechanism. Schema semantics (column names, units, entity strings, free text) is information that numeric-array models cannot represent. If it carries signal, the relevant comparison is a semantics-aware or language-mediated model against both trees and numeric TFMs [14–17], and whether corpus overlap has been excluded must be settled before performance is compared [21,80].

7. Trustworthiness, Interpretability, and Deployment

TFMs change the deployment problem without removing it. They reduce per-dataset training and tuning and add dependencies on the pretraining data, the context construction, the model version, the access mode, and the benchmark regime. Figure 5 maps five risk surfaces to the failure mode each produces and the disclosure each requires; the disclosures are the items of Figure 4.

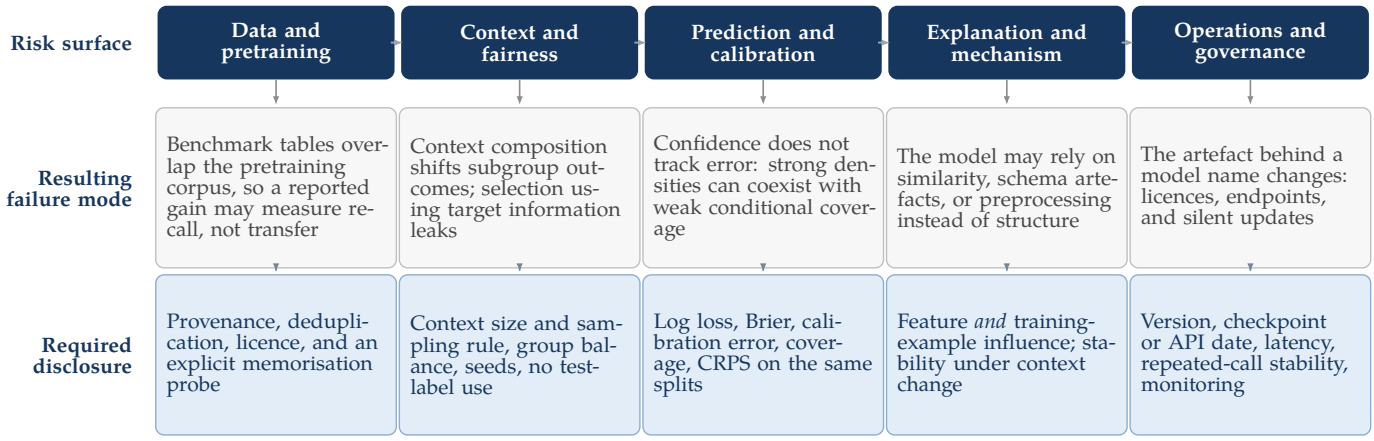


Figure 5. Deployment-readiness map. Each risk surface produces a specific failure mode, and each failure mode implies a specific disclosure.

7.1. Data, Context, and Prediction Risks

Data and pretraining risk depends on how the model was built. Synthetic priors reduce privacy and contamination risk because the pretraining tasks are generated, not copied from public tables [1]. Real and mixed pretraining add realism and require provenance, licensing, and leakage controls [12,13]. The risk is greatest for language-mediated models, whose web or table corpora can overlap with evaluation sources [17,21]. *Context construction* is the second surface. In an in-context model, the selected rows are the effective evidence at prediction time. Fairness work on TabPFN v2, TabICL, and TabDPT shows that context-selection strategies change group-fairness metrics with limited accuracy loss [46], measured against the equality-of-opportunity criterion of Hardt et al. [96]. The same mechanism leaks if context rows are chosen with information unavailable in deployment. *Prediction reliability* is assessed separately from accuracy (Section 5.3), and deployment-like splits change rankings relative to IID evaluation [18]. *Operational readiness* depends on whether the model can be reproduced, served, monitored, and audited. Learned contexts and retrieval-based localisation improve adaptation but change the evaluated workflow [65,66]. Open weights support local auditing, privacy control, ablation, and mechanistic analysis, whereas API-only systems and releases with restrictive licences are harder to replicate when versions, endpoints, or access conditions change.

7.2. Interpretability Across Paradigms

Interpretability is the dimension on which the paradigms differ most, and the difference is structural. Every per-dataset paradigm has a fitted model whose parameters encode what was learned, and explanation interrogates that object. A pure in-context TFM has no such object: its weights encode a learning procedure shared across all tables, and everything specific to the problem lives in the context. What has to be explained therefore splits into the prediction, the context rows that produced it, and the prior that shaped the procedure.

Four distinctions organise the interpretability column of Table 14. The first is intrinsic versus post hoc explanation. Depth-limited trees and additive models are intrinsically interpretable, and additive models with pairwise terms recover much of the accuracy of an unrestricted model while staying readable [89]. Tree ensembles are opaque in aggregate but admit exact Shapley attribution in polynomial time [91], which for an arbitrary model can only be approximated [97]. Deep tabular networks, tabular-to-image pipelines, and TFMs are post hoc by default, and intrinsically interpretable TFM designs remain open in the eligible literature. On the image side, hybrid networks combine Grad-CAM saliency on the image branch with attribution on the raw-feature branch [35].

The second distinction is feature versus example attribution. Feature attribution says which columns mattered; example attribution says which context rows mattered. Only the first is routinely reported, yet for an in-context learner the second identifies the evidence actually used, and it changes with the context-selection rule [46]. Mechanistic analyses ask what the network retrieves or transforms from context [45]. The third is local against global. A global explanation of a tree ensemble is well defined because the fitted model is fixed, whereas for a TFM only the prior and the procedure can be characterised globally, the target of mechanistic probing [45]. The fourth is explanation quality, covering faithfulness, cost, and stability, which no eligible study yet measures for TFM explanations. Stability is the open item: resampling, reordering, and rebalancing the context move the prediction [46], so a TFM explanation should be reported with the context protocol that produced it. With correlated features, whether an attribution describes the fitted model or the data distribution matters more for in-context learners, whose behaviour on a table is a function of that table's rows, so the intended target should be stated. In scientific and engineering applications, corroborating data-driven attributions against an independent physical model is a stronger standard than agreement between attribution methods; a recent engineering study reads explainable-AI attributions against multiphysical simulation to uncover process–structure relationships [98].

Tree ensembles therefore keep a real, and in regulated deployment often decisive, advantage: exact attribution, an auditable fitted structure, and a size–transparency trade-off the practitioner controls by limiting depth. TFMs give this up and offer a form of explanation the other families lack, direct inspection of which training examples support a prediction.

7.3. Deployment Footprint

The families differ more in deployment profile than in accuracy (Table 15), and "training-free" collapses three cost components. The *fit cost* is what it takes to obtain a model for a table. For trees, deep networks, and tabular-to-image pipelines, it is per-dataset training and tuning; for a TFM, it is an amortised pretraining the user does not pay for, plus a context-preparation step the user does. The *per-prediction cost* is independent of the training-set size for a tree ensemble, which evaluates a fixed sequence of comparisons, but scales with the training table for a pure in-context learner, which must hold the context at prediction time. The *artefact footprint* is everything that must be stored, shipped, versioned, and audited.

Three findings follow. The tree advantage on constrained targets is structural: artefact size and per-prediction cost are set by user-chosen parameters and need no accelerator. The TFM constraint is the context, not the weights: the networks are small by current standards, while inference memory grows with the training-table context, to one million rows under 50 GB of GPU memory in the largest reported envelope [10]. Under tight constraints, the natural production route distils the TFM into a compact student model, a route no eligible study yet evaluates. The deployed artefact under hard constraints therefore belongs today to the classical families, and the open question is whether a distilled model beats a directly tuned compact model of the same footprint (Direction D7). No eligible benchmark reports accuracy under a fixed memory or energy budget on a constrained device, which is why Table 15 separates structural properties from measured results.

Table 15. Deployment profile of the main families. Structural entries follow from how each family computes a prediction; measured entries carry their source.

Dimension	Tree Ensembles	Deep and Image-Based Networks	Tabular Foundation Models
Locus of learning	Per dataset, on site	Per dataset, on site	Once, off site; the site supplies context
Deployed artefact	Trees sized by the user	Weights; fitted image layout	Pretrained network plus context rows
Inference hardware	CPU	CPU; GPU for large models	GPU in the reported evaluations [1,10]
Memory at inference	Constant	Constant	Grows with the training-table context; 1M rows under 50 GB of GPU memory [10]
Measured cost	Reference comparator	Layout fit up to 10^4 s [33]	2.8 s against a four-hour search [1]; 5.5 × clinical runtime [19]; order-of-magnitude faster spatial prediction than a hierarchical Bayesian model, costlier per-variable imputation [99]
Size–accuracy control	Direct	Indirect	Not exposed by the pretrained artefact
On-site update	Routine	Routine	Learned-context and retrieval adaptation [65,66]; open-weight releases support local updates
Governance	Local	Local	Varies by release and licence [1,10]

7.4. Evidence Ladder and Practitioner Guidance

Deployment maturity should be stated conservatively, and the *TFM Evidence Ladder* (M0–M4, Figure 6) does so. A model occupies a rung only if every item beneath it is available. The ladder measures the availability of evidence, not its favourability, so unfavourable regime evidence counts and a rung name describes what evidence exists without recommending the model. Its trust items are checkable conditions: subgroup performance means metrics reported per protected or domain-relevant subgroup, and drift-sensitive validation means evaluation under a temporal or otherwise shifted split of the deployment data. Rung M2 requires stratified eligible results on all five regime axes (temporal, grouped, scale, dimensionality, text), because a model's regime profile is unknown until every axis is tested. Stratification on a subset of axes is recorded per axis (Section 9.2) without occupying the rung. Section 9.2 applies the ladder to four models, each the one with the largest body of corpus evidence in its lineage branch: the pure in-context reference model, the scalable in-context line, the language-mediated line, and the structured-task extensions.

A practical validation path starts from the deployment regime, not from the model name, and takes three decisions in order. The split is random only when the intended use is approximately IID, and otherwise grouped, temporal, site-held-out, or external. The baselines include a tuned tree ensemble and, where relevant, a tuned neural or AutoML system at a stated budget. The context protocol states which rows are available at prediction time, how they are sampled, whether sampling uses protected or target-related information, and whether it would be permitted in deployment. Table 16 condenses this guidance by regime.

Table 16. Practitioner decision guide by deployment regime.

Situation	First Choice	Required Check
Small or medium IID data, low tuning budget	Open TFM	Default and tuned GBDT comparison; proper scores whenever the output is used as a probability (high-stakes row)
Few samples, very many features	TFM, tuned GBDT, image hybrid	Transformation cost reported
Temporal or grouped data	Tuned GBDT or deep model	TFM tested on the matching split
Very high dimensionality or cardinality	Model within its envelope	Supported feature limits
Large row count	Scalable TFM	Matched budget against tuned GBDT and AutoML

Table 16. *Cont.*

Situation	First Choice	Required Check
Text-rich tables	Semantics-aware or language model	Contamination and provenance
High-stakes probabilities	Recalibrated model	Proper scores, not accuracy alone
Private or regulated data	Open-weight local model	Data provenance
API-only model	Pinned version	Access date and repeated-call stability
Strict latency or memory	Distilled compact model	Directly tuned model of equal size; on-device cost

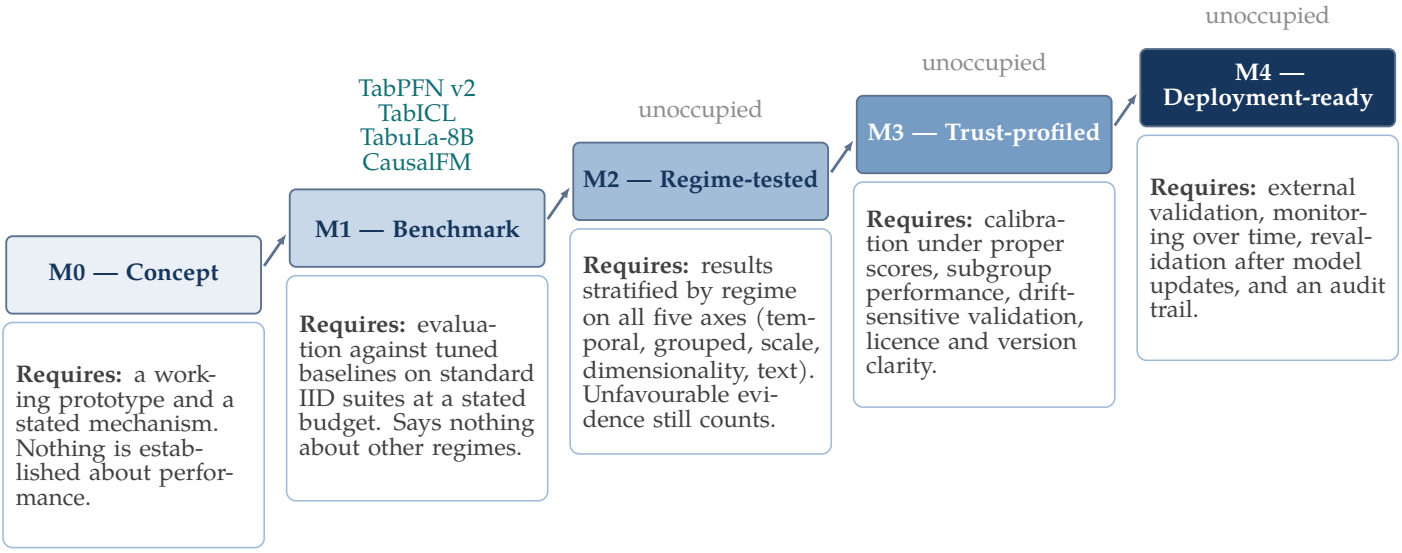


Figure 6. The TFM Evidence Ladder (M0–M4). A model occupies a rung only when every item beneath it is available; the occupancy shown follows the assessment of Section 9.2.

8. Application Evidence by Regime

Application studies are read here for what only applied work can show: whether the advantage survives external validation, how evidence quality differs across domains, and which operational constraints bind in practice. Table 17 grades the domains with comparison-bearing evidence and states the validation that would raise each grade.

Three observations follow. Independence changes the verdict. Clinical prediction is the one domain with a dedicated independent benchmark, and it gives the most conditional result, parity with established methods at higher computational cost [19]. Elsewhere, the studies are largely single-team, so favourable results are demonstrations rather than validations [29,99]. Applied use is broader than validated applied evidence: the corpus keeps five application studies because only those report a comparison capable of testing a claim. And the binding constraint is often not accuracy. It is GPU dependence and runtime in clinical deployment [19], temporal drift and regulatory defensibility in industry [18], and device resources in industrial and IoT intrusion detection, where the published designs build ensembles or hybrids around the model's sample and class ceilings [29,30,100]. In scientific small-data settings, it is the trustworthiness of the predictive distribution [78,99]. The security studies report strong results under random splits on standard intrusion corpora, while temporal splits and adversarial perturbation, the conditions the domain imposes, remain untested. Across domains, success is predicted by the data regime, not the field.

Table 17. Application evidence by domain, graded with the rubric of Table 4, and the validation required to strengthen each claim.

Application Area	Governing Regime	Evidence	Grade (Total: A/B/C/D/E)	Validation Needed
Clinical risk [19,53]	Moderate to large cohorts; strong local baselines	Most AUROC gaps within ±0.02; best baseline beaten in 16.7% of tasks; 5.5× runtime; survival-aware adaptation for censored outcomes	Moderate 7: 2/2/1/2/0	Site-held-out tests; minority strata; calibration; censoring-aware metrics
Industrial and IoT intrusion [29,30,100]	Few incidents; device limits; adversarial drift	Competitive detection through ensembles built around sample and class limits	Moderate 6: 2/2/1/0/1	Temporal and adversarial tests; minority-attack recall; on-device profiling
Geoscience and soil [99]	Small, costly samples	Beats a hierarchical Bayesian baseline on both tasks; order-of-magnitude faster spatial prediction; costlier per-variable imputation	Moderate 5: 1/2/1/1/0	Blocked spatial validation; kriging baselines
Structured-task extensions [28,58,60]	Non-tabular problems posed as tables	Working forecasting, optimisation, and decision workflows	Preliminary 4: 2/0/2/0/0	Backtesting; regret curves

Each grade applies the rubric of Table 4 to the claim that a TFM matches or beats tuned established baselines under the governing regime; A–E are the rubric dimensions.

9. Discussion and Research Agenda

The preceding sections treat each evidence domain on its own. This section reads them against one another and turns the result into an agenda.

9.1. Regime Dependence and Evaluation Broadening

Across the corpus, the factor most consistently associated with whether a TFM matches or exceeds its comparators is the evaluation regime, not the architecture, the prior, or the release date. Read alone, the evaluations of Table 9 seem to disagree; read together, they sample different regions of one regime space and agree once the regime is held fixed. The spatial-encoding benchmark extends the agreement to tabular-to-image pipelines, which meet the same small-N, high-d boundary [33].

A reported result is therefore a statement about a region of that space, and its transferability is bounded by the region's edges. Four boundaries recur. A model validated to a row and feature envelope carries no evidence beyond it. The envelopes differ across the lineage: roughly 10,000 rows and 500 features for TabPFN v2, around 500,000 rows for TabICL under its own settings, and a million rows for TabICLv2 [1,9,10]. Random splits on data with temporal or grouped structure can reorder rankings entirely [18]. A frozen TFM against a lightly tuned tree answers a different question from the same TFM against a four-hour AutoML ensemble. And discrimination and reliability can move in opposite directions in the same model [77].

Reading the domains together also exposes a regularity no single study shows: in the three broadening comparisons of Table 18, extending the split design, the metric family, or the evaluator's independence moved the estimated advantage towards parity. The comparisons differ in directness. The metric and evaluator comparisons are direct, observing the same models under a broader protocol. The split comparison is indirect, because the temporal benchmark evaluates per-dataset deep models and GBDTs, not TFMs [18]. It shows that temporal splits reorder rankings among established models, not that a measured TFM advantage shrank. The comparisons come from different studies, metrics, and settings, not from an exhaustive enumeration, so the regularity is a hypothesis. The corpus also holds counter-examples: TabICL leads TabPFN v2 and CatBoost on the TALENT datasets above 10,000 samples [9], and the drift-adapted TabPFN variant reports a favourable developer result under temporal shift [76]. The small-data results remain real and replicated. The

hypothesis supports a default reading: a favourable result from a narrow regime is an upper bound on the advantage available outside it. This places the burden of demonstration on the broader claim.

Table 18. Three comparisons underlying the evaluation-broadening pattern.

Broadened Axis	from and to	Effect on the TFM Advantage	Evidence
Split design	Random to temporal splits [18]	Rankings reorder; MLPs and GBDTs strongest	Indirect: no TFM evaluated
Metric family	Discrimination to conditional coverage [77]	Coverage trails GBDTs	Direct
Evaluator	Developers to independent clinical benchmark [19]	Parity at 5.5× runtime	Direct

9.2. Disclosure and Evidence Maturity

Reading the trustworthiness risks of Section 7 against the evaluation and application evidence yields one cross-cutting conclusion: the assumptions that carry a TFM result are the ones least often reported. Context protocols, compute accounting, pretraining provenance, and artefact identity each determine what a number means. Each has been shown to move it: context composition shifts group fairness at near-constant accuracy [46], total cost of prediction can invert an apparent efficiency advantage [19], and corpus overlap can substitute recall for generalisation [21,80]. The audit of Table 11 shows how far practice falls short: 3 of 29 comparison-bearing studies evaluate under a deployment-matched split, none of them an independent evaluation of a TFM, and 13 of the 24 that evaluate a pretrained model report no contamination or leakage check. A result without its context rule, total cost of prediction, or contamination checks is less interpretable, not only less reproducible. Evidence maturity is therefore not evidence quantity: a model with many favourable results can sit lower on the Evidence Ladder than one with fewer, if those results rest on unstated assumptions.

Applied to four models that span the families, the ladder separates them by the type of evidence available, not by headline performance. It exposes a structural gap: all four occupy M1 (Figure 6), because M2 requires eligible results stratified on all five regime axes and no eligible study provides them. Single-axis stratification exists, from TabICL's size-stratified TALENT results [9] and the clinical benchmark's cohorts of up to 139,528 patients [19], and Table 19 records it per axis without rung occupancy. Several trust items above M2 are available out of rung order for the pure in-context models: calibration and coverage evidence for TabPFN v2 [77,78], subgroup evidence for TabPFN v2 and TabICL [46], and independent external clinical evidence for TabPFN v2, including unfavourable findings [19]. Drift-sensitive validation exists for none of the four; the drift-adapted variant's temporal-shift evidence [76] bears on H4 but does not validate the frozen reference model. TabuLa-8B is held at M1 by the lack of contamination-controlled transfer evaluation in light of the memorisation evidence [17,21,80], and CausalFM by the lack of independent causal replication under a shared protocol with explicit tests of identification assumptions [54]. No reviewed model reaches M4, which requires post-deployment evidence: performance monitored through drift, revalidation after model or endpoint updates, and an audit trail linking decisions to the artefact that produced them. That vacancy is the lifecycle gap behind Direction D6.

Table 19. Evidence Ladder criterion matrix for the four representative models.

Criterion	TabPFN v2	TabICL	TabuLa-8B	CausalFM
M0—Concept				
Prototype and stated mechanism	[1]	[9]	[17]	[54]
M1—Benchmark				
Tuned baselines, standard suites, stated budget	[1]: 4 h ensemble, AutoML	[9]: TALENT tuned baselines	[17]: XGBoost, TabPFN	[54]: causal estimators
M2—Regime-tested (all five axes required)				
Scale	[19], independent, to 139,528 patients	[9], developer size-stratified	–	–
Temporal	– (drift-adapted variant only [76])	–	–	–
Grouped	–	–	–	–
Dimensionality	– (adapted beyond the envelope [68])	–	–	–
Text	–	–	– (contamination-controlled evaluation outstanding [21,80])	–
M3—Trust-profiled				
Calibration under proper scores	[77,78], incl. unfavourable	–	–	–
Subgroup performance	[46]	[46]	–	–
Drift-sensitive validation	– (variant only [76])	–	–	–
Licence and version clarity	[1]	[9]	[17]	–
M4—Deployment-ready				
External domain validation	[19]	–	–	–
Monitoring, revalidation, audit trail	–	–	–	–
Occupied rung	**M1**	**M1**	**M1**	**M1**

A citation names the corpus source satisfying the item; a parenthesised entry does not count towards occupancy; – no corpus source. A model occupies the highest rung whose items, and all items beneath, are satisfied.

9.3. Research Directions

Seven directions follow from the findings, each naming the change in practice it requires.

D1: Make the regime the unit of reporting, and run the missing head-to-head. Results should be stratified by sample size, dimensionality, split structure, feature type, and cardinality instead of being collapsed into a leaderboard position; the infrastructure exists [20,49]. Any claim of superiority should name both the regime in which it was measured and the regime in which it was not tested. The one comparison the literature still lacks pits TFMs against tuned GBDTs and parametric tabular-to-image hybrids on the same small-N and high-d datasets under one tuning protocol. Both families claim that regime, and neither benchmark includes the other [20,33].

D2: Report discrimination and reliability under one protocol. Probabilistic quality should be evaluated on the same splits, context protocol, and compute budget as accuracy, with proper scoring rules and coverage diagnostics [82,83]. Because the two axes can move in opposite directions [77,78], reporting them separately hides the trade-off where a practitioner needs it. Where recalibration is applied, its cost and its effect on both axes belong in the result.

D3: Standardise and disclose the context and cost protocol. A shared minimum description of what the model was given and what the prediction cost (context size and sampling rule, group balance, ordering, preprocessing, schema exposure, adaptation state, hardware, and total cost of prediction on the evaluated workload) would make

the literature on context selection, fairness, and adaptation aggregable [46,65,66]. Figure 4 states such a minimum.

D4: Build contamination-controlled evaluation for real-data and language-mediated models. Most comparison-bearing studies of a pretrained model report no contamination or leakage check (Table 11). Models pretrained on real or web-scale corpora need evaluation resources whose overlap with the pretraining distribution is known: provenance records, collections built after the pretraining cut-off, and explicit memorisation probes of the kind that have already changed conclusions [21,80]. Synthetic-prior models can answer the same question constructively by characterising the coverage of the generated distribution [1].

D5: Extend the advantage beyond IID through priors and adaptation. Three routes appear in the reviewed work: priors that encode temporal and grouped structure instead of exchangeable rows, of which Drift-Resilient TabPFN is the first eligible instance [76]; retrieval and context construction that select evidence appropriate to a shifted target [12,66]; and adaptation to a deployment distribution at a stated cost [65, 68]. The order-aware prior experiment of H4 (Section 6.2) decides whether the non-IID deficit is an objective mismatch or a missing re-fitting step; its temporal half now has a developer-reported answer, which leaves the group-aware prior and an independent replication open. Each route should be evaluated against the regime it targets, not the IID suites on which the paradigm was established.

D6: Move to mechanistic and lifecycle-aware validation. Two kinds of evidence are missing at opposite ends of the pipeline. The field lacks a predictive account of when in-context tabular inference is sound. The Bayesian and mechanistic readings are complementary but incomplete, and the ablations proposed for H2 and H3 would turn regime dependence from a finding into a design tool [27,45]. And it lacks evidence generated after deployment: performance tracked across drift, revalidation after model or endpoint updates, and audit trails linking a decision to the artefact that produced it. Until the latter exists, the top rung of the Evidence Ladder stays unoccupied by construction.

D7: Evaluate under deployment budgets, and settle the distillation question. Two changes would make the comparison of Section 7.3 empirical instead of structural. The first is budgeted evaluation: accuracy reported at fixed memory, latency, and energy ceilings on the class of device the application targets; the infrastructure for varying the inference budget already exists in the leading benchmarks [20]. The second is the comparison the distillation route implies: whether a compact model distilled from a TFM outperforms a directly tuned compact model of the same footprint. No eligible study reports such a comparison; an open implementation would settle the question.

9.4. Limitations

Four limitations bound what this review can establish, and each has a mitigation. First, one author carried out screening, extraction, and scoring, so the grades are author-assigned assessments rather than consensus judgements. The mitigations are those of Section 2.3: printed dimension scores, stated coding rules, recomputable counts, and the sensitivity analysis of Table 13. Second, identification- and screening-stage counts were not retained, so the selection is documented as a structure rather than as a PRISMA-style flow [26], and the routes are post hoc title classifications that cannot demonstrate the completeness of the search. A mechanical classification rule any reader can rerun, the corpus composition of Table 2, and the account of the evidence set aside under the peer-review criterion stand in their place. Third, the eligibility criterion excludes preprint-only evidence, which in this field includes some of the broadest regime-stratified results. Section 5.1 states what

admitting it would change, and the affected analyses are re-examined whenever such a work reaches a reviewed venue. Fourth, the eligible comparisons lean towards classification and developer-affiliated evaluations, and the tabular-to-image coverage is concentrated in one research group. Affiliations and dependency groups are recorded per source, separated by the independence dimension, and counted once by the replication dimension.

10. Conclusions

This review asked where tabular foundation models hold a supported advantage over established tabular learners. One criterion, reusable cross-table pretraining, separates them from the per-dataset paradigms they compete with, and five interacting design axes decide which comparisons are fair (Q1). Across the comparisons that meet this condition, the data regime decides more than the architecture (Q2). TFMs match or exceed tuned tree ensembles on small-to-medium IID tables when little tuning budget is available, whereas tuned trees and deep networks keep temporal, feature-engineered data. Larger tables are covered by developer evaluations and one independent clinical study. Grouped splits and frozen use beyond the feature envelope remain untested, so no claim can be made there in either direction. Tabular-to-image pipelines obey the same small-sample, high-dimensional boundary and have not been compared with TFMs directly. Among the graded claims (Q3), only the small-data claim is Strong. Training-free use is supported for accuracy but contested for subgroup outcomes until context protocols are disclosed, the sufficiency of synthetic priors is moderate, and the claims on predictive distributions, structured-task extensions, language-model table predictors, and image encodings remain Preliminary. The evidence a high-stakes deployment needs (Q4), a matched split, proper scores, a disclosed context, and an account of cost, is seldom reported, and no reviewed model has post-deployment evidence. In each broadening comparison assembled here, a wider evaluation narrowed the estimated advantage. This remains a working hypothesis with stated counter-examples. The Reporting Checklist and the Evidence Ladder let a study state what its result shows and where it applies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics15215156/s1, Table S1: The reviewed corpus, with the evidence-matrix codes of the comparison-bearing studies; Table S2: Per-source scores behind the claim grades (Table 12) and domain grades (Table 17) of the article; Table S3: Records set aside under the peer-review criterion: preprints and reports that meet a content criterion but have no peer-reviewed version. References [1,2,7–25,27–35,44–80,84–88,92,93,99,100] are cited in the Supplementary Materials.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were generated. All data analysed were extracted from the cited published works, and the counts and grades derived from them are reported in this article.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	area under the receiver operating characteristic curve
AutoML	automated machine learning
CRPS	continuous ranked probability score
GBDT	gradient-boosted decision tree
ICL	in-context learning
IID	independent and identically distributed
LLM	large language model
MLP	multilayer perceptron
PFN	prior-data fitted network
SCM	structural causal model
TFM	tabular foundation model

References

1. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
2. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 7499–7519. [CrossRef] [PubMed]
3. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
4. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
5. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018.
6. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 507–520.
7. McElfresh, D.; Khandagale, S.; Valverde, J.; Prasad C, V.; Ramakrishnan, G.; Goldblum, M.; White, C. When do neural nets outperform boosted trees on tabular data? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 76336–76369.
8. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
9. Qu, J.; Holzmüller, D.; Varoquaux, G.; Le Morvan, M. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2025; Volume 267, pp. 50817–50847.
10. Qu, J.; Holzmüller, D.; Varoquaux, G.; Le Morvan, M. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2026; Volume 306, pp. 102477–102555.
11. Zeng, Y.; Dinh, T.; Kang, W.; Mueller, A.C. TabFlex: Scaling tabular learning to millions with linear attention. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2025; Volume 267, pp. 74051–74079.
12. Ma, J.; Thomas, V.; Hosseinzadeh, R.; Labach, A.; Cresswell, J.C.; Golestan, K.; Yu, G.; Caterini, A.L.; Volkovs, M. TabDPT: Scaling tabular foundation models on real data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 172692–172722.
13. Zhang, X.; Maddix Robinson, D.; Yin, J.; Erickson, N.; Ansari, A.F.; Han, B.; Zhang, S.; Akoglu, L.; Faloutsos, C.; Mahoney, M.W.; et al. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 15795–15840.
14. Arazi, A.; Shapira, E.; Reichart, R. TabSTAR: A tabular foundation model for tabular data with text fields. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 172108–172161.

15. Spinaci, M.; Polewczyk, M.; Schambach, M.; Thelin, S. ConTextTab: A semantics-aware tabular in-context learner. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 147005–147040.
16. Kim, M.J.; Grinsztajn, L.; Varoquaux, G. CARTE: Pretraining and transfer for tabular learning. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 23843–23866.
17. Gardner, J.; Perdomo, J.C.; Schmidt, L. Large scale transfer learning for tabular data via language modeling. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 45155–45205.
18. Rubachev, I.; Kartashev, N.; Gorishniy, Y.; Babenko, A. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025; pp. 35166–35202.
19. Shaktah, L.A.; Gustav, M.; Lenz, T.; Liang, J.; Hilgers, L.; Carrero, Z.I.; Kather, J.N. Established machine learning matches tabular foundation models in clinical predictions. *BMC Med. Inform. Decis. Mak.* **2026**, *26*, 364. [CrossRef]
20. Erickson, N.; Purucker, L.; Tschalzev, A.; Holzmüller, D.; Mutalik Desai, P.; Salinas, D.; Hutter, F. TabArena: A living benchmark for machine learning on tabular data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, San Diego, CA, USA, 2–7 December 2025; Volume 38.
21. Gorla, A.; Puduppully, R. The illusion of generalization in tabular language models. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2026; Volume 306, pp. 36369–36396.
22. Müller, S.; Hollmann, N.; Pineda Arango, S.; Grabocka, J.; Hutter, F. Transformers can do Bayesian inference. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
23. Bazgir, O.; Zhang, R.; Dhruba, S.R.; Rahman, R.; Ghosh, S.; Pal, R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nat. Commun.* **2020**, *11*, 4391. [CrossRef] [PubMed]
24. Sun, B.; Yang, L.; Zhang, W.; Lin, M.; Dong, P.; Young, C.; Dong, J. SuperTML: Two-dimensional word embedding for the precognition on structured tabular data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 2973–2981. [CrossRef]
25. Iqbal, M.I.; Mukta, M.S.H.; Hasan, A.R.; Islam, S. A dynamic weighted tabular method for convolutional neural networks. *IEEE Access* **2022**, *10*, 134183–134198. [CrossRef]
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
27. Nagler, T. Statistical foundations of prior-data fitted networks. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 25660–25676.
28. Hoo, S.B.; Müller, S.; Salinas, D.; Hutter, F. From tables to time: Extending TabPFN-v2 to time series forecasting. *Trans. Mach. Learn. Res.* **2026**. Available online: https://jmlr.org/tmlr/papers/bib/KIkQj8VOUY.bib (accessed on 26 September 2026).
29. Ruiz-Villafranca, S.; Roldán-Gómez, J.; Castelo Gómez, J.M.; Carrillo-Mondéjar, J.; Martinez, J.L. A TabPFN-based intrusion detection system for the Industrial Internet of Things. *J. Supercomput.* **2024**, *80*, 20080–20117. [CrossRef]
30. Ruiz-Villafranca, S.; Roldán-Gómez, J.; Carrillo-Mondéjar, J.; Martinez, J.L.; Gañán, C.H. WFE-Tab: Overcoming limitations of TabPFN in IIoT-MEC environments with a weighted fusion ensemble-TabPFN model for improved IDS performance. *Future Gener. Comput. Syst.* **2025**, *166*, 107707. [CrossRef]
31. Castillo-Cara, M.; Talla-Chumpitaz, R.; García-Castro, R.; Orozco-Barbosa, L. TINTO: Converting tidy data into image for classification with 2-dimensional convolutional neural networks. *SoftwareX* **2023**, *22*, 101391. [CrossRef]
32. Liu, J.; González-Fernández, D.; Castillo-Cara, M.; García-Castro, R. TINTOlib: A Python library for transforming tabular data into synthetic images for deep neural networks. *SoftwareX* **2025**, *32*, 102444. [CrossRef]
33. Liu, J.; Castillo-Cara, M.; García-Castro, R. A comprehensive benchmark of spatial encoding methods for tabular data with deep neural networks. *Inf. Fusion* **2026**, *130*, 104088. [CrossRef]
34. Castillo-Cara, M.; Martínez-Gómez, J.; Ballesteros-Jerez, J.; García-Varea, I.; García-Castro, R.; Orozco-Barbosa, L. MIMO-based indoor localisation with hybrid neural networks: Leveraging synthetic images from tidy data for enhanced deep learning. *IEEE J. Sel. Top. Signal Process.* **2025**, *19*, 559–571. [CrossRef]
35. Mondragon-Ruiz, G.; Liu, J.; Castillo-Cara, M.; García-Castro, R. Interpretable CNN–KAN hybrid architectures for tabular data with synthetic image encoding. *Inf. Process. Manag.* **2026**, *63*, 104954. [CrossRef]
36. Arik, S.Ö.; Pfister, T. TabNet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 6679–6687. [CrossRef]
37. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; Volume 34, pp. 18932–18943.

38. Gorishniy, Y.; Rubachev, I.; Kartashev, N.; Shlenskii, D.; Kotelnikov, A.; Babenko, A. TabR: Tabular deep learning meets nearest neighbors. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024; pp. 18209–18249.
39. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025; pp. 77899–77935.
40. Kadra, A.; Lindauer, M.; Hutter, F.; Grabocka, J. Well-tuned simple nets excel on tabular datasets. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; Volume 34, pp. 23928–23941.
41. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
42. Holzmüller, D.; Grinsztajn, L.; Steinwart, I. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 26577–26658.
43. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020; Volume 33, pp. 1877–1901.
44. Stoian, M.C.; Giunchiglia, E.; Lukasiewicz, T. A survey on deep learning approaches for tabular data generation: Utility, alignment, fidelity, privacy, diversity, and beyond. *Trans. Mach. Learn. Res.* **2026**. [CrossRef]
45. Ye, H.J.; Liu, S.Y.; Chao, W.L. A closer look at TabPFN v2: Understanding its strengths and extending its capabilities. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 135605–135637.
46. Kenfack, P.; Kahou, S.E.; Aïvodji, U. Towards fair in-context learning with tabular foundation models. *Trans. Mach. Learn. Res.* **2026**. [CrossRef]
47. Somvanshi, S.; Das, S.; Javed, S.A.; Antariksa, G.; Hossain, A. A survey on tabular data: From tree-based methods to tabular deep learning. *ACM Comput. Surv.* **2026**, *58*, 426. [CrossRef]
48. Jiang, J.P.; Liu, S.Y.; Cai, H.R.; Zhou, Q.L.; Ye, H.J. Representation learning for tabular data: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6488–6508. [CrossRef]
49. Liu, S.Y.; Cai, H.R.; Zhou, Q.L.; Yin, H.H.; Zhou, T.; Jiang, J.P.; Ye, H.J. TALENT: A tabular analytics and learning toolbox. *J. Mach. Learn. Res.* **2025**, *26*, 1–16.
50. Sun, Q.; Yuan, H.; Huang, Y.; Zhang, Z.; Fu, X.; Wang, R.; Zhou, H.; Li, J.; Wu, J.; Yu, P.S. A survey on foundation models for structured data: Tabular, time series, and graphs. *IEEE Trans. Knowl. Data Eng.* **2026**, *38*, 6302–6320. [CrossRef]
51. van Breugel, B.; van der Schaar, M. Position: Why tabular foundation models should be a research priority. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 48976–48993.
52. Mueller, A.C.; Curino, C.; Ramakrishnan, R. MotherNet: Fast training and inference via hyper-network transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025; pp. 76666–76686.
53. Pham, M.K.; Cotugno, L.; Sîrbu, A.; Mai, T.T.; Crane, M.; Bezbradica, M. Tabular foundation models for clinical survival analysis via survival-aware adaptation. In Proceedings of the International Conference on Artificial Intelligence in Healthcare (AIiH), Lecture Notes in Computer Science, London, UK, 26–28 August 2026; Volume 16875, pp. 315–328. [CrossRef]
54. Ma, Y.; Frauen, D.; Javurek, E.; Feuerriegel, S. Foundation models for causal inference via prior-data fitted networks. In Proceedings of the International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, 23–27 April 2026; pp. 79065–79098.
55. Balazadeh, V.; Kamkari, H.; Thomas, V.; Li, B.; Ma, J.; Cresswell, J.C.; Krishnan, R.G. CausalPFN: Amortized causal effect estimation via in-context learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 154945–154984.
56. Robertson, J.; Reuter, A.; Guo, S.; Hollmann, N.; Hutter, F.; Schölkopf, B. Do-PFN: In-context learning for causal effect estimation. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 174811–174848.
57. Dooley, S.; Khurana, G.S.; Mohapatra, C.; Naidu, S.V.; White, C. ForecastPFN: Synthetically-trained zero-shot forecasting. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 2403–2426.
58. Schiff, D.; Lindenbaum, O.; Efroni, Y. ICR-RL: Deep reinforcement learning via in-context-regression. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2026; Volume 306, pp. 108204–108217.
59. Müller, S.; Feurer, M.; Hollmann, N.; Hutter, F. PFNs4BO: In-context learning for Bayesian optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 25444–25470.
60. Yu, R.T.Y.; Picard, C.; Ahmed, F. GIT-BO: High-dimensional Bayesian optimization with tabular foundation models. In Proceedings of the International Conference on Learning Representations (ICLR), Rio de Janeiro, Brazil, 23–27 April 2026; pp. 111636–111663.

61. Eremeev, D.; Platonov, O.; Bazhenov, G.; Babenko, A.; Prokhorenkova, L. GraphPFN: A prior-data fitted graph foundation model. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2026; Volume 306, pp. 28149–28165.
62. Lourenço, A.; Gama, J.; Xing, E.P.; Marreiros, G. In-context learning of evolving data streams with tabular foundational models. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Jeju, Republic of Korea, 9–13 August 2026; pp. 973–982. [CrossRef]
63. da Costa Merlin, G.; Furtado Silva, D. TS2TabPFN: Time series classification and extrinsic regression through feature extraction and a tabular foundation model. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2026; Volume 16947, pp. 555–572. [CrossRef]
64. Kim, M.J.; Lefebvre, F.; Brison, G.; Perez-Lebel, A.; Varoquaux, G. Table foundation models: On knowledge pre-training for tabular learning. *Trans. Mach. Learn. Res.* **2025**. [CrossRef]
65. Feuer, B.; Schirrmeister, R.T.; Cherepanova, V.; Hegde, C.; Hutter, F.; Goldblum, M.; Cohen, N.; White, C. TuneTables: Context optimization for scalable prior-data fitted networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 83430–83464.
66. Thomas, V.; Ma, J.; Hosseinzadeh, R.; Golestan, K.; Yu, G.; Volkovs, M.; Caterini, A. Retrieval & fine-tuning for in-context tabular models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 108439–108467.
67. Liu, S.Y.; Ye, H.J. TabSwift: An efficient tabular foundation model with row-wise attention. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2026; Volume 306, pp. 79463–79479.
68. Liu, S.Y.; Ye, H.J. TabPFN unleashed: A scalable and effective solution to tabular classification problems. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2025; Volume 267, pp. 40043–40068.
69. Zhu, B.; Shi, X.; Erickson, N.; Li, M.; Karypis, G.; Shoaran, M. XTab: Cross-table pretraining for tabular transformers. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 43181–43204.
70. Wang, Z.; Sun, J. TransTab: Learning transferable tabular transformers across tables. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 2902–2915.
71. Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; Sontag, D. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*; PMLR: Cambridge, MA, USA, 2023; Volume 206, pp. 5549–5581.
72. Yan, J.; Zheng, B.; Xu, H.; Zhu, Y.; Chen, D.Z.; Sun, J.; Wu, J.; Chen, J. Making pre-trained language models great on tabular prediction. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024; pp. 23570–23593.
73. Wang, Z.; Gao, C.; Xiao, C.; Sun, J. MediTab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Jeju, Republic of Korea, 3–9 August 2024; pp. 6062–6070. [CrossRef]
74. Bonet, D.; Mas Montserrat, D.; Giró-i-Nieto, X.; Ioannidis, A.G. HyperFast: Instant classification for tabular data. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 11114–11123. [CrossRef]
75. Arbel, M.; Salinas, D.; Hutter, F. EquiTabPFN: A target-permutation equivariant prior fitted network. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), San Diego, CA, USA, 2–7 December 2025; Volume 38, pp. 62586–62609.
76. Helli, K.; Schnurr, D.; Hollmann, N.; Müller, S.; Hutter, F. Drift-Resilient TabPFN: In-context learning temporal distribution shifts on tabular data. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 98742–98781.
77. de Melo Costa, J.L.; Popineau, F.; Rimmel, A.; Doan, B.L. High performance, low reliability: Uncertainty benchmarking for tabular foundation models. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Bruges, Belgium, 22–24 April 2026; pp. 115–120. [CrossRef]
78. Johnson, T.R.; Ben-Jacob, K.; Negarandeh, N.; Vendrell-Gallart, O.; Bostanabad, R. On the uncertainty quantification ability of tabular foundation models. *Comput. Sci. Eng.* **2026**, *28*, 75–85. [CrossRef]
79. Vanschoren, J.; van Rijn, J.N.; Bischl, B.; Torgo, L. OpenML: Networked science in machine learning. *ACM SIGKDD Explor. Newsl.* **2014**, *15*, 49–60. [CrossRef]
80. Bordt, S.; Nori, H.; Rodrigues, V.; Nushi, B.; Caruana, R. Elephants never forget: Memorization and learning of tabular data in large language models. In Proceedings of the Conference on Language Modeling (COLM), Philadelphia, PA, USA, 7–9 October 2024.
81. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. [CrossRef]

82. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
83. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
84. Sharma, A.; Vans, E.; Shigemizu, D.; Boroevich, K.A.; Tsunoda, T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci. Rep.* **2019**, *9*, 11399. [CrossRef] [PubMed]
85. Zhu, Y.; Brettin, T.; Xia, F.; Partin, A.; Shukla, M.; Yoo, H.; Evrard, Y.A.; Doroshow, J.H.; Stevens, R.L. Converting tabular data into images for deep learning with convolutional neural networks. *Sci. Rep.* **2021**, *11*, 11325. [CrossRef] [PubMed]
86. Matsuda, T.; Uchida, K.; Saito, S.; Shirakawa, S. HACNet: End-to-end learning of interpretable table-to-image converter and convolutional neural network. *Knowl.-Based Syst.* **2024**, *284*, 111293. [CrossRef]
87. Alenizy, H.A.; Berri, J. Transforming tabular data into images via enhanced spatial relationships for CNN processing. *Sci. Rep.* **2025**, *15*, 17004. [CrossRef]
88. Medeiros Neto, L.; da Silva Neto, S.R.; Endo, P.T. A comparative analysis of converters of tabular data into image for the classification of arboviruses using convolutional neural networks. *PLoS ONE* **2023**, *18*, e0295598. [CrossRef]
89. Lou, Y.; Caruana, R.; Gehrke, J.; Hooker, G. Accurate intelligible models with pairwise interactions. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 623–631. [CrossRef]
90. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
91. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
92. Damri, A.; Last, M.; Cohen, N. Towards efficient image-based representation of tabular data. *Neural Comput. Appl.* **2024**, *36*, 1023–1043. [CrossRef]
93. Lara-Abelenda, F.J.; Chushig-Muzo, D.; Peiro-Corbacho, P.; Gómez-Martínez, V.; Wägner, A.M.; Granja, C.; Soguero-Ruiz, C. Transfer learning for a tabular-to-image approach: A case study for cardiovascular disease prediction. *J. Biomed. Inform.* **2025**, *165*, 104821. [CrossRef]
94. Hager, P.; Menten, M.J.; Rueckert, D. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23924–23935. [CrossRef]
95. Du, S.; Zheng, S.; Wang, Y.; Bai, W.; O'Regan, D.P.; Qin, C. TIP: Tabular-image pre-training for multimodal classification with incomplete data. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 15073, pp. 478–496. [CrossRef]
96. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016; Volume 29.
97. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 4765–4774.
98. Wu, M.; Yao, Z.; Ye, L.; Ge, J.; Zha, J.; Zhu, Y. A synergistic framework combining explainable AI and multiphysical simulation for uncovering process–structure relationships in mask electrolyte jet machining. *Adv. Eng. Inform.* **2026**, *76*, 105089. [CrossRef]
99. Saito, T.; Otake, Y.; Wu, S. Applying a tabular foundation model to geotechnical site characterization. *Geod. AI* **2026**, *6*, 100040. [CrossRef]
100. Li, Z.; Xiong, C.; Zheng, K.; Li, Q. An RF-TabPFN-based framework for few-shot IoT network attack recognition using LASSO-RFE feature selection. *IEEE Access* **2025**, *13*, 151452–151465. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.