Learning the expansion basis instead of choosing it: a gated orthogonal-polynomial functional link network for diabetes risk classification

Revisiting Legendre and Chebyshev functional link architectures under a modern evaluation protocol

[1]Chandrasekhar Panda [2]Saswat Swain
[1]Ph.D Scholar, [2]Ph.D Scholar,
[1,2]KSOM - KIIT Deemed to be University , India
[1]contact@chandrasekharpanda.com [2]saswat.mck@gmail.com

Abstract—Single layer functional link networks remain attractive for clinical risk screening because they train in milliseconds, expose every coefficient to inspection and run on hardware that cannot host a deep model. Their accuracy, however, hinges on a design decision usually made by hand and defended after the fact, namely which family of orthogonal functions should expand the input. Comparative studies of Legendre and Chebyshev expansions, including the one that motivated this work, have reached conflicting conclusions. We remove the decision from the designer and hand it to the learner. The proposed adaptive gated orthogonal polynomial functional link network expands every attribute simultaneously in four function families and learns a temperature annealed gate deciding, separately for each attribute, how those families are mixed. Capacity is controlled by a shrinkage term growing with expansion order, motivated by the decay of orthogonal expansion coefficients of smooth functions, and by a group sparsity term applied through a proximal step. Legendre, Chebyshev and trigonometric functional link networks are recovered exactly as degenerate gates, so the architecture strictly generalises the models it replaces. Under nested cross validation on three public diabetes cohorts, against nine baselines including random forests and two gradient boosting libraries, the proposed network attains the highest Matthews correlation on the Pima cohort and joins the leading statistical group overall while using one to three orders of magnitude fewer parameters. We further show that the fixed threshold, accuracy oriented protocol of earlier functional link studies badly overstates the clinical usefulness of these classifiers on imbalanced cohorts. *(Abstract)*

Index Terms—Functional link network, orthogonal polynomial expansion, basis gating, diabetes risk prediction, group sparsity, calibration. *(key words)*

I. Introduction

Diabetes mellitus is now among the most heavily modelled conditions in clinical informatics, and the reason is arithmetic rather than fashion: the disease is common [1], its onset is preceded by a long and largely silent phase, and the variables that carry most of the predictive signal are cheap to obtain. A risk score computed from eight routine measurements is therefore worth building even when its discrimination is modest, because the alternative in most primary care settings is no score at all.

A large part of the classical literature on this problem was written with functional link architectures. A functional link network expands each scalar input through a fixed family of nonlinear functions and then fits a single linear layer on the expanded vector. Because there is no hidden layer, training reduces to a convex or near convex problem, the parameter count stays in the tens, and every coefficient can be traced back to a named clinical variable and a named basis function. Those properties matter for deployment on point of care devices and for the kind of audit that a clinical decision support tool has to survive.

The awkward part of this family of models is the expansion itself. Trigonometric expansions were proposed first, Chebyshev expansions were introduced to reduce the cost of evaluating them, and Legendre expansions were introduced in turn on the grounds that they are cheaper still. Each substitution has been reported as an improvement, sometimes on the same benchmark. Our own earlier comparison of a Legendre network against a Chebyshev functional link network on the Pima cohort concluded in favour of the Legendre expansion by roughly one percentage point of test accuracy on a single split; re-running that comparison here under repeated cross validation leaves the two architectures statistically indistinguishable. The conclusion we draw is not that the earlier result was wrong but that the question was badly posed. Which polynomial family suits an attribute is a property of that attribute, not of the dataset, and it should not be settled by a global search over four alternatives.

This paper proposes an architecture in which the choice is made per attribute and made by gradient descent. Every input is expanded concurrently in four families, Legendre, Chebyshev of the first kind, probabilists' Hermite and

trigonometric, and a softmax gate over families is learned jointly with the output weights. The gate temperature is annealed so that the mixture starts soft, which keeps the early gradients informative, and hardens towards a near categorical selection, which keeps the fitted model readable. Two further mechanisms keep the expanded model from overfitting the small cohorts on which it will be used: an order graded ridge penalty whose weight grows with the polynomial degree, and a group sparsity penalty over the coefficient block belonging to each attribute, applied through a proximal step rather than by subgradient descent.

The contributions are four. An architecture that subsumes the Legendre, Chebyshev and trigonometric functional link networks as special cases, so that any accuracy gained is gained over a strict superset rather than over a differently tuned competitor. A training rule with closed form gate gradients and a proximal operator for the sparsity term, implementable without an automatic differentiation framework. An evaluation deliberately harsher than the earlier literature: nested cross validation, nine competing learners, selection confined to inner folds, and reporting centred on Matthews correlation and calibration. And a quantified demonstration that the older protocol is not merely imprecise but systematically optimistic on imbalanced cohorts.

II. Related work

A. Functional link architectures and learnable bases

The functional link idea originates with Pao, who replaced hidden layers by a fixed nonlinear expansion followed by a single adaptive layer [4]. Chebyshev expansions were adopted for nonlinear system identification because their three term recurrence is cheap and numerically stable [5], and the argument was transferred to classification in data mining [3]. Legendre expansions were proposed as a further simplification, and comparative studies on channel equalisation and medical classification followed [6]. Later work concentrated on the training rule, replacing least mean squares by swarm or evolutionary search, or on capacity, adding competitive units to obtain universal approximation [11]. The expansion family itself remained a fixed design choice throughout.

Interest in learning the univariate nonlinearity has revived through Kolmogorov Arnold networks, which place a learnable spline on every edge instead of a fixed activation on every node [7]. The variants that followed replaced the spline by Chebyshev polynomials [8], by wavelets [9] and by other polynomial groups [10], and a recent survey catalogues several dozen such substitutions [12]. The pattern is instructive: the community has repeated, at greater expense, the substitution exercise that the functional link literature performed two decades earlier. A single layer functional link network with a learned per attribute basis occupies the intersection of the two lines of work and inherits the cost profile of the older one.

B. Tabular classification and evaluation practice

Systematic benchmarks continue to find that tree ensembles are hard to beat on medium sized tabular problems, and that reported neural advantages often dissolve once hyper parameter search is equalised [13], [14], [25]; those benchmarks also fix the protocol we adopt. The measurement literature has separately argued that accuracy is a poor summary under class imbalance and that the Matthews correlation coefficient should be preferred [17], and that classifiers are frequently miscalibrated in ways no discrimination metric can reveal [18]. The functional link studies on diabetes data predate both arguments and report accuracy at a fixed threshold, the practice we quantify and reject in Section V-C.

III. Proposed architecture

Let $x \in \mathbb{R}^d$ denote an attribute vector and $y \in \{0, 1\}$ the diabetes label. Each attribute is mapped to the interval $[-1, 1]$ by an affine transform whose limits are the first and ninety ninth training percentiles, with clipping outside that range; using percentiles rather than extrema prevents a single outlying insulin reading from compressing the entire expansion into a narrow band.

A. Multi-family functional expansion

Four families are evaluated for every attribute. The three polynomial families are generated by their standard three term recurrences,

$$(n+1)\,P_{n+1} = (2n+1)\,x\,P_n - n\,P_{n-1};\quad T_{n+1} = 2x\,T_n - T_{n-1};\quad He_{n+1} = x\,He_n - n\,He_{n-1}, \tag{1}$$

for the Legendre, Chebyshev of the first kind and probabilists' Hermite families respectively, and the fourth family reproduces the original functional link expansion,

$$\varphi_n(x) = \sin(k\pi x) \text{ for odd } n,\ \cos(k\pi x) \text{ for even } n,\ k = \lceil n/2 \rceil. \tag{2}$$

All four are initialised with the zeroth term absorbed into the bias and the first term equal to x. Because the four families produce columns of very different scale, and because a softmax gate can only mix quantities that are commensurable, each expanded column is standardised using training fold statistics,

$$\hat{\varphi}_{i,b,n} = (\varphi_{i,b,n} - \mu_{i,b,n}) / \sigma_{i,b,n}. \tag{3}$$

We refer to this step as basis whitening; the ablation in Section V-D shows that omitting it costs accuracy on the smallest cohort.

B. Adaptive basis gating

For attribute i a vector of logits $g_i \in \mathbb{R}^B$ is learned and converted to mixing weights by a tempered softmax,

$$\alpha_{i,b} = \exp(g_{i,b} / \tau) / \Sigma_{b'} \exp(g_{i,b'} / \tau), \tag{4}$$

and the expanded features seen by the output layer are the corresponding mixtures,

$$\psi_{i,n}(x) = \Sigma_b\, \alpha_{i,b}\, \hat{\varphi}_{i,b,n}(x_i). \tag{5}$$

The temperature follows a geometric schedule from $\tau_0 = 2$ to $\tau_1 = 0.25$. A high initial temperature keeps all four families active while the output weights are still uninformative, preventing the gate from committing to whichever family happens to receive the larger random coefficients; the low final temperature drives the mixture towards one family per attribute, so the fitted model reduces to a list of attribute basis pairs.

Pairwise products of the scaled attributes form an optional second block. Products are ranked by the absolute correlation between the product and the residual of a linear probe fitted on the training fold, and only the highest ranked are retained. The decision output is

$$z(x) = w_0 + \Sigma_i \Sigma_n w_{i,n} \psi_{i,n}(x) + \Sigma_p v_p q_p(x), \quad \hat{y} = \sigma(z), \tag{6}$$

with σ the logistic function and q the retained products.

C. Objective

Training minimises a cost sensitive log likelihood together with two structured penalties,

$$J = \sum_m c_\gamma \ell(\hat{y}_m, y_m) / M + \lambda^G \sum_i \sqrt{N} \|w_i\|_2 + \lambda_o \sum_{i,n} n^r w_{i,n}^2 + \lambda_v \|v\|_1, \tag{7}$$

where ℓ is binary cross entropy and the class weights c are inversely proportional to class frequency. The second term is a group lasso [24] over the coefficient block of each attribute and therefore performs attribute selection rather than coefficient selection. The third term is the mechanism we call order graded shrinkage.

Proposition 1 (order graded shrinkage). If the conditional log odds is a univariate function of Sobolev smoothness r on [−1, 1], its coefficients in the three polynomial families decay at rate $O(n^{-r})$. A quadratic penalty with weight n^γ is the maximum a posteriori estimator under a zero mean Gaussian prior whose variance decays at the same rate, so γ encodes an assumed smoothness rather than an arbitrary strength. We fix γ = 2 and treat only λ_o as tunable.

A practical consequence is that the expansion order need not be selected by cross validation over integers: high order terms survive only where the data pay for them.

D. Learning rule

Write $\delta_m = c_\gamma (\hat{y}_m - y_m) / M$. Differentiating (6) and (7) gives the output layer gradient

$$\partial J / \partial w_{i,n} = \sum_m \delta_m \psi_{i,n}(x_m) + 2 \lambda_o n^r w_{i,n}, \tag{8}$$

and, applying the softmax Jacobian to (4), the gate gradient

$$\partial J / \partial g_{i,b} = \alpha_{i,b} (a_{i,b} - \sum_{b'} \alpha_{i,b'} a_{i,b'}) / \tau, \quad a_{i,b} = \sum_m \delta_m \sum_n w_{i,n} \hat{\varphi}_{i,b,n}(x_m). \tag{9}$$

Both are evaluated in closed form; no automatic differentiation is required. The smooth part of the objective is optimised by adaptive moment estimation [19], after which the non smooth terms are applied exactly by their proximal operators [20], block soft thresholding for the group lasso and scalar soft thresholding for the interaction weights,

$$w_i \leftarrow w_i \max(0, 1 - \eta \lambda^G \sqrt{N} / \|w_i\|_2), \quad v_p \leftarrow \operatorname{sgn}(v_p) \max(0, |v_p| - \eta \lambda_v). \tag{10}$$

Algorithm 1 summarises one pass. Early stopping monitors inner validation loss; the epoch budget so obtained is then reused when the model is refitted on the whole training partition.

Algorithm 1 Training AGOP-FLN

Input: training set, order N, families B, penalties λ^G, λ_o, λ_v, epochs E
scale attributes to [−1, 1] by training percentiles; build and whiten φ
screen and retain the top interaction products
initialise $w \sim N(0, 0.05^2)$, $g \leftarrow 0$, $w_0 \leftarrow$ logit of prevalence
for e = 1 to E do
  $\tau \leftarrow \tau_0 (\tau_1/\tau_0)^{(e/E)}$ ▷ anneal
  for each mini-batch: $\alpha \leftarrow \operatorname{softmax}(g/\tau)$, $\psi \leftarrow \Sigma^b \alpha \hat{\varphi}$, $\hat{y} \leftarrow \sigma(z)$;
    Adam update of w, g, v, w_0 using (8) and (9);
    apply the proximal steps (10)
  evaluate inner validation loss; keep the best state
end for
Output: w, g, v, w_0 and the fitted gate α

E. Relationship to earlier architectures and cost

Proposition 2 (strict generalisation). Fixing $\alpha_{i,b} = 1$ for a single family b and 0 elsewhere, setting $\lambda^G = \lambda_o = \lambda_v = 0$, disabling whitening and replacing the logistic loss by squared error recovers exactly the Legendre network, the Chebyshev functional link network or the trigonometric functional link network according to the choice of b. Every such model is therefore a point in the parameter space explored here, and the gate can only fail to match the best of them by an optimisation error, not by a representational one.

Evaluating the expansion costs O(dBN) operations per sample and the output layer a further O(dN + |S|); memory is O(dBN). With d = 8, B = 4, N = 2 and no interaction block, the fitted Pima model carries 53 free parameters against 25 for the legacy networks, 2,531 for the multilayer perceptron and 31,330 tree nodes for the random forest. The gate contributes dB additional parameters that are discarded after training if a hard selection is made.

IV. Experimental design

A. Cohorts

Three public cohorts are used, summarised in Table I. The Pima Indians cohort is the benchmark on which the functional link literature on diabetes was built and is retained for continuity. The Sylhet cohort is almost entirely binary symptom responses and therefore probes the regime in which a polynomial expansion has little to work with. The Behavioral Risk Factor Surveillance System extract supplies a large, strongly imbalanced screening problem of the kind a deployed risk score would face; a stratified subsample of thirty thousand records preserving the original prevalence was drawn to keep the nested protocol affordable.

On the Pima cohort, zero entries in plasma glucose, diastolic pressure, triceps thickness, serum insulin and body mass index are physiologically impossible and are treated as missing; medians computed on the training partition only are imputed. This departs from the earlier studies, which used the raw zeros, and is applied identically to every learner.

B. Protocol

The outer estimator is repeated stratified k fold cross validation, five folds repeated three times for the small cohorts and three folds for the large one. Within each outer training partition a stratified three fold inner loop produces out of fold predictions used for three purposes: selecting hyper parameters by inner area under the curve, selecting the decision threshold by inner Matthews correlation, and fixing the epoch budget of the proposed model. Every learner is then refitted on the entire outer training partition. No test fold participates in any selection, and all learners observe identical data and identical selection machinery.

Nine baselines are used: regularised logistic regression, a radial basis support vector machine, a multilayer perceptron, a random forest, XGBoost [15], LightGBM [16], and faithful reimplementations of the Legendre, Chebyshev and trigonometric functional link networks trained by the least mean squares rule of the original studies with third order expansions and a step size of 0.08. The support vector machine was not run on the large cohort because its training cost is quadratic in sample size. Each baseline receives its own inner grid of comparable size.

Reported measures are accuracy, sensitivity, specificity, F measure, area under the receiver operating characteristic curve, Matthews correlation, Brier score and expected calibration error over ten bins. Significance is assessed by the Wilcoxon signed rank test over folds with Holm correction, by Cliff's delta as an effect size, and across cohorts by the Friedman test with the Nemenyi critical difference [21].

TABLE I
COHORTS USED IN THE STUDY

Cohort	Records	Attr.	Positive	Character
Pima Indians [2]	768	8	34.9 %	continuous
Sylhet [22]	520	16	61.5 %	mostly binary
BRFSS 2015 [23]	30 000	21	13.9 %	mixed

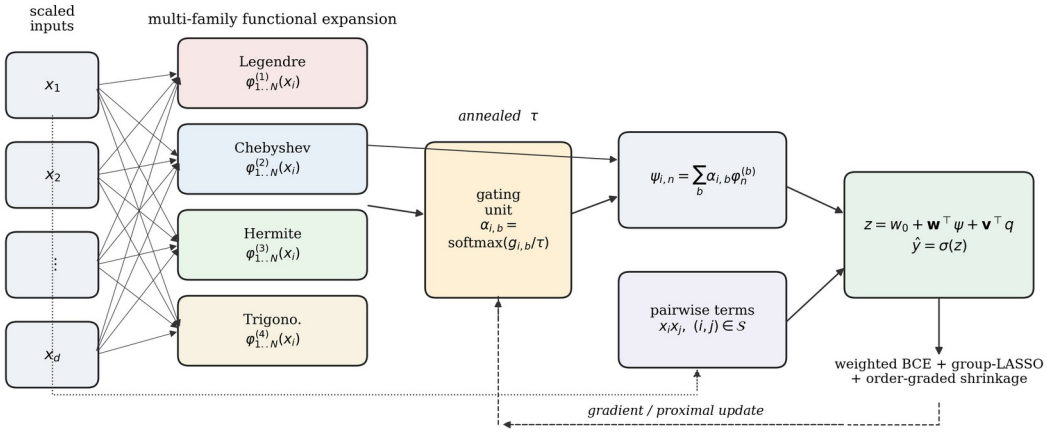

Fig. 1. Adaptive gated orthogonal-polynomial functional link network. Each attribute is expanded concurrently in four function families; a tempered softmax gate mixes them per attribute, and a single linear layer combines the mixed expansions with screened pairwise products.

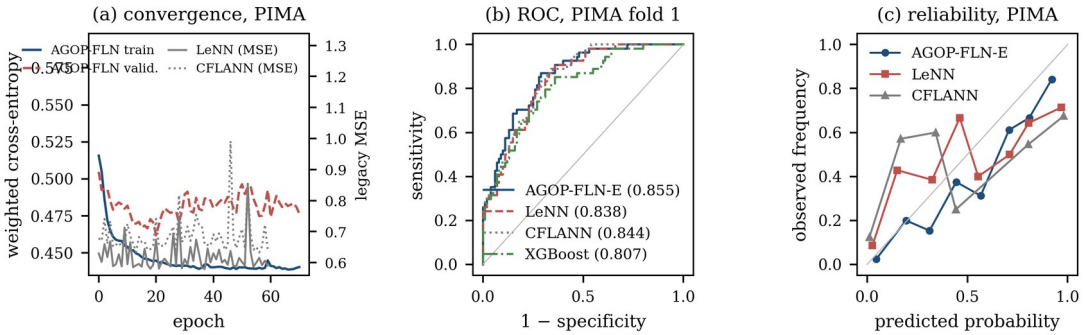

Fig. 2. (a) Convergence of the proposed model against the least-mean-squares training of the legacy networks on the Pima cohort; note the different vertical scales. (b) Receiver operating characteristic on the first test fold. (c) Reliability curves: the proposed model tracks the diagonal, the legacy networks do not.

V. Results and discussion

A. Classification performance

Table II reports the outer fold means. On the Pima cohort the proposed network is first on every listed measure, reaching 0.769 accuracy, 0.784 sensitivity, 0.852 area under the curve and 0.527 Matthews correlation, against 0.498 for the random forest and 0.461 and 0.469 for the Legendre and Chebyshev predecessors. The margin over the legacy functional link models is roughly six points of Matthews correlation, obtained with a model of comparable size and the same absence of hidden layers.

On the Sylhet cohort the ordering changes. The proposed model improves on every functional link predecessor by seven to ten points of Matthews correlation and on logistic regression by ten, but the random forest is genuinely better (0.969 against 0.931). This is expected rather than anomalous: fifteen of the sixteen attributes are binary, a polynomial evaluated at two points spans at most a two dimensional space, and axis aligned splits are the natural hypothesis class here.

On the large imbalanced cohort the five strongest learners are separated by less than four thousandths of Matthews correlation, with the ensembled variant of the proposed model nominally first at 0.368 and an area under the curve of 0.826. The legacy networks are far behind, at 0.654 and 0.654 area under the curve, which is the clearest evidence in this study that a fixed third order expansion trained by least mean squares does not scale to twenty one attributes.

The ensembled variant does not behave uniformly. Averaging nine differently seeded models changes Matthews correlation by -0.0014 on Pima, which is not significant, and by 0.0065 on the large cohort, where variance reduction has something to act on. Since the ensemble multiplies the parameter count by nine, we regard the single model as the deployable configuration.

B. Statistical assessment

The significance markers in Table II summarise the paired tests against the proposed model. On the Pima cohort the improvement is significant after Holm correction against logistic regression, the multilayer perceptron, both

gradient boosting libraries and all three functional link predecessors, with Cliff's delta between 0.46 and 0.65, and is not significant against the random forest. On the Sylhet cohort the proposed model is significantly better than the three predecessors and logistic regression and significantly worse than the random forest. On the large cohort the three available folds leave the signed rank test without resolving power, its smallest attainable p value being 0.25, so only effect sizes are meaningful there and they favour the proposed model against the three predecessors with a delta of 1.

Pooling the ten learners that were run on all three cohorts over 33 cohort fold blocks, the Friedman test rejects the hypothesis of equal ranks decisively ($\chi^2 = 108.2$, $p \approx 3 \times 10^{-19}$). The critical difference at the five percent level is 2.36 ranks. Fig. 4(a) shows the resulting picture: the random forest (3.14), the ensembled proposal (3.24) and the single proposal (3.39) form the leading group and cannot be separated from one another, while all three legacy functional link networks, logistic regression and the multilayer perceptron fall outside the critical difference of the best rank.

C. What the older protocol concealed

The comparison that motivated this work, and most of the functional link literature on diabetes, reports accuracy at a fixed decision threshold of one half. Applying both protocols to the same fitted legacy models leaves the two balanced cohorts essentially unchanged, the largest shift in Matthews correlation being 0.017. On the imbalanced screening cohort, Table III, the picture is different.

Under the older protocol the Legendre network reports 0.861 accuracy, a figure that would read as a strong result in an abstract, while detecting 10.8 percent of the diabetic cases. Selecting the threshold on inner folds moves the same model to 0.825 accuracy and 39.9 percent sensitivity, raising Matthews correlation from 0.184 to 0.289. The trigonometric variant behaves the same way, trading eight points of accuracy for forty three of sensitivity. An accuracy oriented protocol therefore does not merely add noise; it rewards the classifier that abstains from the positive class, precisely the failure mode a screening instrument must avoid.

TABLE II
OUTER-FOLD MEAN PERFORMANCE. SN IS SENSITIVITY; BEST VALUE PER COHORT AND METRIC IN BOLD. A DAGGER MARKS A MATTHEWS CORRELATION SIGNIFICANTLY BELOW, AND A DOUBLE DAGGER SIGNIFICANTLY ABOVE, THAT OF AGOP-FLN (WILCOXON SIGNED-RANK, HOLM-CORRECTED, 0.05). THE LAST COLUMN IS THE AVERAGE FRIEDMAN RANK OVER ALL COHORT-FOLD BLOCKS. THE SUPPORT VECTOR MACHINE WAS NOT RUN ON BRFSS.

	Pima Indians				Sylhet				BRFSS 2015				Rank
Model	**Acc**	**Sn**	**AUC**	**MCC**	**Acc**	**Sn**	**AUC**	**MCC**	**Acc**	**Sn**	**AUC**	**MCC**	**avg.**
Logistic regression	0.736	0.689	0.834	0.447†	0.916	0.923	0.972	0.827†	0.786	0.647	0.822	0.362	7.32
SVM (RBF)	0.734	0.797	0.830	0.481	0.974	0.983	0.997	0.945	—	—	—	—	—
MLP	0.734	0.571	0.803	0.410†	0.864	0.859	0.939	0.729†	0.762	**0.700**	0.822	0.361	7.53
Random forest	0.751	0.780	0.836	0.498	**0.985**	**0.990**	**0.999**	**0.969‡**	0.795	0.619	0.823	0.361	3.14
XGBoost [15]	0.723	0.786	0.817	0.458†	0.972	0.969	0.995	0.941	0.765	0.695	0.825	0.363	4.68
LightGBM [16]	0.719	0.752	0.802	0.441†	0.976	0.978	0.995	0.950	0.767	0.697	0.823	0.365	4.89
LeNN [6]	0.739	0.704	0.840	0.461†	0.931	0.926	0.966	0.859†	0.825	0.399	0.654	0.289	6.79
CFLANN [5]	0.728	0.742	0.839	0.469†	0.926	0.920	0.968	0.849†	**0.840**	0.340	0.654	0.284	6.88
Trig. FLANN [4]	0.719	0.739	0.834	0.454†	0.920	0.935	0.967	0.832†	0.790	0.568	0.713	0.329	7.14
AGOP-FLN (proposed)	**0.769**	0.784	**0.852**	**0.527**	0.967	0.974	0.988	0.931	0.777	0.667	0.820	0.361	3.39
AGOP-FLN-E (proposed)	0.765	**0.797**	0.851	0.526	0.968	0.976	0.988	0.933	0.791	0.640	**0.826**	**0.368**	3.24

TABLE III
EFFECT OF THE EVALUATION PROTOCOL ON THE LEGACY NETWORKS, BRFSS COHORT. “2015” FIXES THE THRESHOLD AT 0.5; “NEW” SELECTS IT ON INNER FOLDS. THE SAME FITTED MODELS ARE USED IN BOTH COLUMNS.

Cohort	**Model**	**Acc 2015**	**Acc new**	**Sn 2015**	**Sn new**	**MCC 2015**	**MCC new**
BRFSS	LeNN	0.861	0.825	0.108	0.399	0.184	0.289
BRFSS	CFLANN	0.863	0.840	0.117	0.340	0.201	0.284
BRFSS	FLANN-trig	0.865	0.790	0.140	0.567	0.232	0.329

Calibration tells a consistent story. Expected calibration error for the legacy networks on the Pima cohort is 0.150 and 0.204, against 0.113 for the proposed model. The reason is structural: a network trained by least mean squares on ±1 targets and read through a hyperbolic tangent produces a score with no probabilistic meaning, whereas the proposed model optimises a proper scoring rule. Fig. 2(c) shows the reliability curves.

D. Ablation

Fig. 4(b) reports the change in Matthews correlation when each mechanism is removed, averaged over folds. Cost sensitive weighting is the largest single contributor, worth 0.050 on average and 0.107 on the imbalanced cohort alone. Interaction terms come second, worth 0.104 on the symptom cohort where several attributes are only informative jointly, and nothing at all on the other two, which is why the screening step retains a variable number of products.

The gating result is the interesting one and it is not the result we expected. Replacing the learned gate by a uniform mixture costs 0.0118 on average, so learning the mixture is better than averaging over families. Replacing it by a single fixed Legendre basis costs only 0.0019, which is negligible. The gate does not, on these cohorts, extract accuracy a well chosen fixed basis could not. What it does is remove the need to choose, at a cost of dB parameters and no measurable accuracy penalty. Since selecting among four families by cross validation multiplies the

training budget fourfold, dissolving the question is a more useful outcome than winning it.

Order graded shrinkage and temperature annealing contribute smaller amounts, concentrated on the Pima cohort (0.025 and 0.012 respectively), consistent with their role as regularisers: they matter where data are scarce and are neutral where they are not.

E. What the fitted model says

Fig. 3(a) shows the gate learned on the Pima cohort. Serum insulin and body mass index select the trigonometric family almost categorically, with weights of 0.90 and 0.80, and plasma glucose and the pregnancy count select it less firmly; triceps thickness, the pedigree function and age leave the gate close to uniform over the three polynomial families, which is what the gate does when an attribute carries little signal and the group penalty has already shrunk its coefficients. Across the cohort the trigonometric family wins 62 percent of attributes and the Hermite family the remainder; the Legendre and Chebyshev families win none. That the family favoured by the original comparison is selected for no attribute is worth stating plainly.

Fig. 3(b) ranks attributes by the norm of their coefficient block, the natural relevance measure under a group penalty. Body mass index, plasma glucose and age dominate in that order and the remaining five are shrunk to a small fraction of their magnitude, recovering the ordering that clinical practice reports without any post hoc attribution step: the ranking is a fitted parameter, not an explanation constructed afterwards.

On the large cohort, Fig. 3(c), the gate is nearly one hot everywhere. Binary indicators overwhelmingly select the trigonometric family, which on two point support reduces to a scaled indicator, whereas the three genuinely ordinal attributes, general health, mental health days and physical health days, select the Legendre, Chebyshev and Hermite families respectively. The gate is therefore separating ordinal attributes from binary ones rather than choosing arbitrarily.

TABLE IV
MODEL SIZE AND SINGLE-CORE INFERENCE LATENCY PER RECORD. P DENOTES THE PIMA COHORT, B THE BRFSS COHORT. TREE ENSEMBLES ARE REPORTED AS NODE COUNTS.

Model	**Par. (P)**	**µs (P)**	**Par. (B)**	**µs (B)**
Logistic reg.	9	3.5	22	0.2
MLP	2,531	3.6	2,593	0.7
Random forest	31,330	183.5	353,315	41.7
XGBoost	2,663	12.2	5,956	3.7
LightGBM	3,867	38.6	6,000	13.2
LeNN	25	1.7	64	0.7
AGOP-FLN	53	3.2	184	8.6

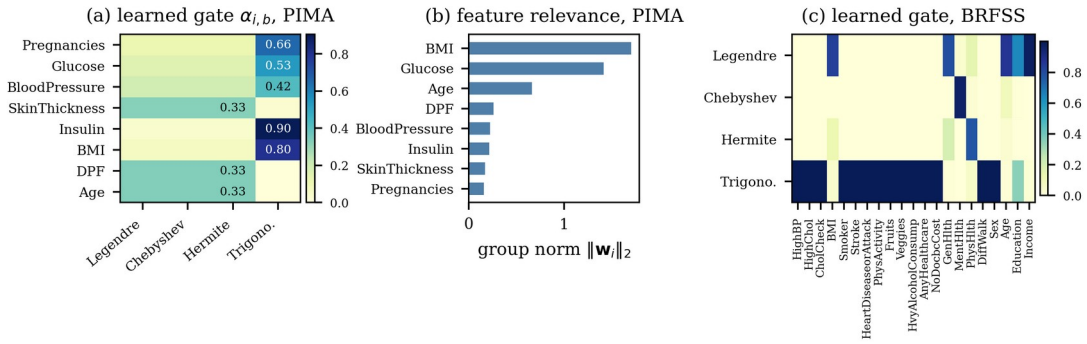

Fig. 3. (a) Gate learned on the Pima cohort; the annotated value is the winning mixing weight. (b) Attribute relevance measured by the norm of each coefficient block. (c) Gate learned on the BRFSS cohort, showing near-categorical selection and a clear separation between binary and ordinal attributes.

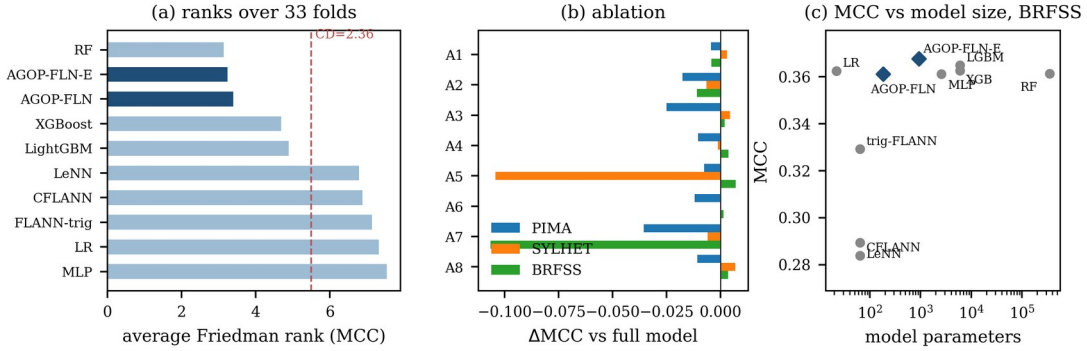

Fig. 4. (a) Average Friedman rank over all cohort-fold blocks; the dashed line marks the Nemenyi critical difference from the best rank. (b) Change in Matthews correlation when each mechanism is removed. (c) Matthews correlation against parameter count on the BRFSS cohort.

F. Cost

Table IV reports size and latency. On the large cohort the fitted model holds 184 parameters against 353,315 nodes for the random forest of equal accuracy, a ratio close to two thousand, and predicts in 8.6 microseconds per record. Training takes 5.5 seconds of single core time, slower than the compiled parallel boosting libraries and faster than the random forest. For an instrument refitted rarely and evaluated constantly, on hardware that may be a clinic microcontroller, that is the right trade.

G. Limitations

Four limitations bound the claims. None of the cohorts is a prospective sample from a deployment population; the Pima cohort in particular is small, single ethnicity and forty years old, and the large cohort is self reported survey data whose labels carry a reporting bias no classifier can correct. Three outer folds on that cohort leave the paired tests underpowered, so the pooled Friedman analysis carries the statistical weight there. Finally, the ranking against tree ensembles is cohort dependent, and we make no claim that a gated polynomial expansion dominates gradient boosting in general: the claim is that it reaches the same statistical group at one to three orders of magnitude less model size, and that it dominates the functional link architectures it was designed to replace.

VI. Conclusion

Whether a Legendre or a Chebyshev expansion is the better basis for diabetes classification has no dataset level answer. Making the basis a per attribute learned quantity removes the question and, on the cohort where the earlier study was performed, produces the best result in this comparison: a six point gain in Matthews correlation over both predecessors and a significant advantage over gradient boosting, using fifty three parameters. Across three cohorts the model joins the leading statistical group with the random forest while remaining two to three orders of magnitude smaller.

The secondary finding is methodological. Evaluating these classifiers by accuracy at a fixed threshold rewards abstention from the positive class and, on an imbalanced screening cohort, reports an eighty six percent accurate model that finds one diabetic case in ten. Future comparisons in this family should report threshold free discrimination, a correlation measure and a calibration measure, and should select thresholds inside the resampling loop.

Two extensions look worthwhile: allowing the gate to mix orders as well as families, and stacking two gated expansions to obtain a shallow Kolmogorov Arnold network with closed form gradients, which would bridge the two literatures reviewed in Section II.

Acknowledgment

The author thanks the curators of the three public repositories from which these cohorts were obtained. The simulation code, the fold level result tables and the scripts that generated every figure are released with this manuscript so that the reported numbers can be reproduced end to end.

References

[1] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: International Diabetes Federation, 2021.

[2] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in Proc. Annu. Symp. Comput. Appl. Med. Care, 1988, pp. 261–265.

[3] B. B. Misra and S. Dehuri, “Functional link artificial neural network for classification task in data mining,” J. Comput. Sci., vol. 3, no. 12, pp. 948–955, 2007.

[4] Y. H. Pao, Adaptive Pattern Recognition and Neural Networks. Reading, MA: Addison-Wesley, 1989.

[5] J. C. Patra and A. C. Kot, “Nonlinear dynamic system identification using Chebyshev functional link artificial neural networks,” IEEE Trans. Syst., Man, Cybern. B, vol. 32, no. 4, pp. 505–511, Aug. 2002.

[6] P. Satpathy, “A comparative study of Legendre neural network and Chebyshev functional link artificial neural network for diabetes data classification,” unpublished.

[7] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, “KAN: Kolmogorov-Arnold networks,” arXiv:2404.19756, 2024.

[8] S. SS, K. AR, G. R, and A. KP, “Chebyshev polynomial-based Kolmogorov-Arnold networks: an efficient architecture for nonlinear function approximation,” arXiv:2405.07200, 2024.

[9] Z. Bozorgasl and H. Chen, “Wav-KAN: wavelet Kolmogorov-Arnold networks,” arXiv:2405.12832, 2024.

[10] S. T. Seydi, “Exploring the potential of polynomial basis functions in Kolmogorov-Arnold networks: a comparative study of different groups of polynomials,” arXiv:2406.02583, 2024.

[11] A. Lotfi and M. R. Akbarzadeh-T, “A competitive functional link artificial neural network as a universal approximator,” Soft Comput., vol. 22, no. 14, pp. 4849–4861, 2018.

[12] S. Somvanshi, S. A. Javed, M. M. Islam, D. Pandit, and S. Das, “A survey on Kolmogorov-Arnold network,” ACM Comput. Surv., vol. 58, no. 2, art. 55, 2025.

[13] L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?” in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 507–520.

[14] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, “Deep neural networks and tabular data: a survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 6, pp. 7499–7519, Jun. 2024.

[15] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[16] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, “LightGBM: a highly efficient gradient boosting decision tree,” in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3146–3154.

[17] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, art. 6, 2020.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning, 2017, pp. 1321–1330.

[19] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in Proc. 3rd Int. Conf. Learning Representations, 2015.

[20] N. Parikh and S. Boyd, “Proximal algorithms,” Found. Trends Optim., vol. 1, no. 3, pp. 127–239, 2014.

[21] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.

[22] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques,” in Computer Vision and Machine Intelligence in Medical Image Analysis, Advances in Intelligent Systems and Computing, vol. 992. Singapore: Springer, 2020, pp. 113–125.

[23] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, “Building risk prediction models for type 2 diabetes using machine learning techniques,” Prev. Chronic Dis., vol. 16, art. 190109, 2019.

[24] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” J. Roy. Statist. Soc. B, vol. 68, no. 1, pp. 49–67, 2006.

[25] R. Shwartz-Ziv and A. Armon, “Tabular data: deep learning is not all you need,” Inf. Fusion, vol. 81, pp. 84–90, 2022.