Natural Resources for Human Health
eISSN: 2583-1194

Original Research

**Received** 15 May 2026
**Revised** 18 June 2026
**Accepted** 08 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1350–1365

**KEYWORDS**:
Chronic kidney disease, attention network, missing data, calibrated probability, clinical decision support, tabular deep learning.

# MIRA-Net: A Missingness-Informed Residual Attention Network for Chronic Kidney Disease Prediction

Pooja Dodamani[1], Juslin F.[2], Navya R.[3]*, Rajesh I S[4], Pallavi R[5], Prathap Paduvalahippe Basavaraju[6], Tripti R. Kulkarni[7]

[1]Assistant Professor, School of Engineering; Department of Computer Science and Engineering, Dayananda Sagar University, India. Email: pooja.s.dodamani-cse@dsu.edu.in
[2]Assistant Professor, Dept. of ECE, Vidyavardhaka College of Engineering, Mysuru, India. Email: juslinfranklin@gmail.com
[3]Assistant Professor, School of Engineering, Department of Electronics and Communication Engineering, Dayananda Sagar University, India. Email: navya-ece@dsu.edu.in
[4]Department of AI&ML, BMSIT&M, Bengaluru, Karnataka, India. Email: rajeshaiml@bmsit.in
[5]Department of CY, ATMECE ,Mysuru, India. Email: pallaviraj830@gmail.com
[6]Associate Professor, Department of Electronics & Communication Engineering, Navkis College of Engineering, Hassan, India. Email: prathap_0293@navkisce.ac.in
7Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India. Email: kulkarnitripti@gmail.com
*Corresponding Author: Navya R. (navya-ece@dsu.edu.in)

**ABSTRACT:** Chronic kidney disease progresses silently, and the routine laboratory panels that could reveal it early are almost never complete: entries are skipped and a clinician is left reasoning from a partial record. Most published models delete such records or fill the gaps with column statistics, discarding the information carried by the pattern of absence itself. This work presents MIRA-Net, a missingness-informed residual attention network that treats every clinical variable as a token and encodes whether that variable was actually measured. A learned absence prototype replaces the value term for unobserved entries, a reliability gate scales each token by the evidence it carries, and two residual attention blocks fuse the tokens before a temperature-scaled class-token read-out. The network holds 79 618 parameters and classifies one record in 1.42 ms on a single processor core. Two independent cohorts totalling 600 records were used, one with 10.55 per cent of its entries absent. Under repeated stratified ten-fold cross-validation, MIRA-Net reached 99.67 ± 1.07 per cent accuracy, 99.74 per cent F1 score and 99.99 per cent area under the curve on the first cohort, with a Brier score of 0.0032. The reliability gate ranked specific gravity, haemoglobin and red blood cell morphology highest, matching accepted nephrological practice and giving every decision an auditable trace.

## 1. INTRODUCTION

Chronic kidney disease (CKD) affects close to seven hundred million people worldwide and was responsible for roughly 1.2 million deaths in 2017, a figure that has risen faster than for most other non-communicable conditions [1]. Its clinical

character is what makes it dangerous. Kidney function can fall to a third of normal before a patient reports any symptom, so the disease is frequently identified only when replacement therapy is already the sole remaining option [2]. Detection therefore depends on laboratory surveillance rather than on presentation, and the value of an automatic detector lies in how well it exploits the panels that are already ordered in routine care. Machine learning and deep learning have been applied across this surveillance problem in most of its clinical forms, and recent surveys map the resulting body of predictive methods [3].

Those panels are rarely complete. In the cohort studied here, 60.8 per cent of patient records are missing at least one of the twenty-four recorded variables, and 10.55 per cent of all cells are empty. The gaps are not uniform: red blood cell morphology, which requires manual microscopy, is absent far more often than serum creatinine, which is produced by an automated analyser. Absence in a clinical record is thus informative. It reports on which investigations a clinician judged necessary, on which samples were adequate, and on the resources of the site where the patient was seen. A model that deletes incomplete rows loses most of its data; a model that fills the gaps with a column median treats a measured value and a guessed one as equally trustworthy [4], [5].

A second problem receives even less attention. Published CKD detectors are almost always assessed on accuracy alone, and on the standard benchmark that metric has saturated: several unrelated methods now report figures above ninety-eight per cent [6], [7], [8]. Accuracy at that level no longer separates one model from another, and it says nothing about whether the probability attached to a prediction can be trusted. A screening tool that reports ninety per cent confidence should be correct in about nine of ten such cases; if it is not, a clinician cannot use the number to triage, and the model is only a label generator [9], [10]. Calibration, not accuracy, is the axis on which a saturated benchmark still discriminates.

This paper addresses both concerns with a single architecture. The contributions are as follows.

1. A tokenised representation of a clinical record in which each variable contributes one token built from a variable identity embedding, a value projection and, when the entry is unobserved, a learned absence prototype that replaces the value term rather than approximating it.
2. A per-token reliability gate that lets the network attenuate variables whose evidence is weak, so that observed and reconstructed entries do not enter the attention computation on equal terms.
3. A calibrated read-out combining a small snapshot ensemble with single-parameter temperature scaling, reported alongside accuracy through the Brier score and the expected calibration error.
4. An evidence ranking read directly off the reliability gate, requiring no post hoc explainer, and evaluation on two independent cohorts under repeated stratified cross-validation with a paired non-parametric test against eight baselines.

Section 2 reviews related work, Section 3 describes the cohorts, Section 4 develops the model, Section 5 sets out the protocol, Section 6 reports results, Section 7 discusses them, and Section 8 concludes.

## 2. RELATED WORK

**Classical learning on renal panels.** The earliest systematic studies applied standard classifiers to the twenty-four-variable panel released by Rubini et al. [11]. Rady and Anwar compared probabilistic neural networks, multilayer perceptrons, support vector machines and radial basis function networks for staging, finding the probabilistic network strongest but with accuracy falling to 96.9 per cent at the intermediate stages where the decision matters most [12]. Yashfi et al. added lifestyle variables and reported 97.12 per cent from a random forest over twenty selected features [13]. Almasoud and Ward pursued the opposite goal, minimising cost rather than error, and showed that gradient boosting over three inexpensive tests retains most of the discriminative power of the full panel [14]. Chittora et al. conducted the broadest comparison, evaluating seven algorithms with and without feature selection and with synthetic oversampling, and obtained 98.46 per cent from a linear support vector machine on the full feature set [6]. Senan et al. combined recursive feature elimination with four classifiers and reported similar figures [8]. Amaresh and Sundaram went further along the same axis, wrapping swarm optimisation around the feature subset to build a screening framework aimed at early detection [15]. Selection of that kind decides which variables

enter the model, but it still says nothing about the ones that were never measured in the first place. Each of these studies imputes or discards missing entries before learning begins, and none reports a calibration measure.

**Hybrid and deep approaches.** Qin et al. built an integrated classifier that routes a record to a nearest-neighbour or random forest branch according to a confidence criterion, and used it to fill missing values before classification [7]. Chen et al. moved to a deep convolutional model deployed on an Internet of Medical Things platform, gaining deployment flexibility at the cost of a much larger parameter budget [16]. Tomasev et al. demonstrated at a different scale that sequence models over longitudinal renal records can anticipate acute injury days in advance [17], which confirms the clinical value of the direction but requires data of a richness that a single outpatient panel does not offer. Elsewhere in abdominal diagnosis the field has moved to generative and graph-regularised architectures, as in the conditional diffusion model Radhika et al. apply to pancreatic tumour segmentation [18]; those designs exploit spatial structure that a tabular panel simply does not contain, which is why the present work builds on token attention rather than on convolution or diffusion.

**Attention over tabular data.** Outside nephrology, several architectures adapt the transformer to tables. TabTransformer embeds categorical columns and passes them through self-attention while leaving continuous columns untouched [19]. TabNet uses sequential attention to select a sparse feature subset at each decision step, yielding interpretability by construction [20]. FT-Transformer tokenises every column, continuous and categorical alike, and appends a class token, establishing that a plain transformer is competitive with gradient boosting when properly tuned [21]. The tokenisation used here follows that line [22], but none of these designs represents whether a value was observed. On a clinical panel, where the missingness rate reaches 10.55 per cent and is strongly variable-dependent, that omission discards a signal the present work sets out to use.

**Calibration.** Guo et al. showed that modern networks are systematically overconfident and that a single temperature parameter fitted on held-out data corrects most of the distortion without altering the ranking of predictions [9]. Lakshminarayanan et al. established that averaging a small ensemble improves both accuracy and uncertainty estimates [23]. Both mechanisms are used in the read-out described in Section 4.

## 3. MATERIALS

Two independent cohorts were used, summarised in Table 1.

**Table 1.** Composition of the two evaluation cohorts.

| Property | Cohort A | Cohort B |
|---|---|---|
| Records | 400 | 200 |
| CKD positive | 250 | 128 |
| CKD negative | 150 | 72 |
| Variables used | 24 | 24 |
| Numeric variables | 14 | 14 |
| Nominal variables | 10 | 10 |
| Absent entries (%) | 10.55 | 4.17 |
| Records with a gap (%) | 60.8 | 100.0 |
| Value resolution | Continuous | Interval midpoint |

**Cohort A** is the CKD panel collected over two months at a nephrology centre in Tamil Nadu, India, and released through the UCI repository [11]. It contains 400 patient records, 250 with a confirmed CKD diagnosis and 150 without, described by 24 variables: 14 numeric measurements and 10 nominal indicators. Missing entries are marked in the source file and were preserved as such. Across the whole matrix 10.55 per cent of entries are absent and 60.8 per cent of records have at least one gap.

Natural Resources for Human Health
eISSN: 2583-1194

**Cohort B** is an independently curated version of the same clinical protocol in which every continuous variable has been discretised into ordered intervals and the estimated glomerular filtration rate and disease stage are supplied. It contains 200 records, 128 positive and 72 negative. Each interval label was mapped to its midpoint, and open intervals to their finite bound, so that the variable set aligns with cohort A. Diastolic pressure appears in this cohort only as a binary indicator rather than a pressure, so it was marked unobserved throughout and left for the model to handle through its absence mechanism. The stage and filtration rate columns were withheld, since both are derived from the diagnosis and would leak the label.

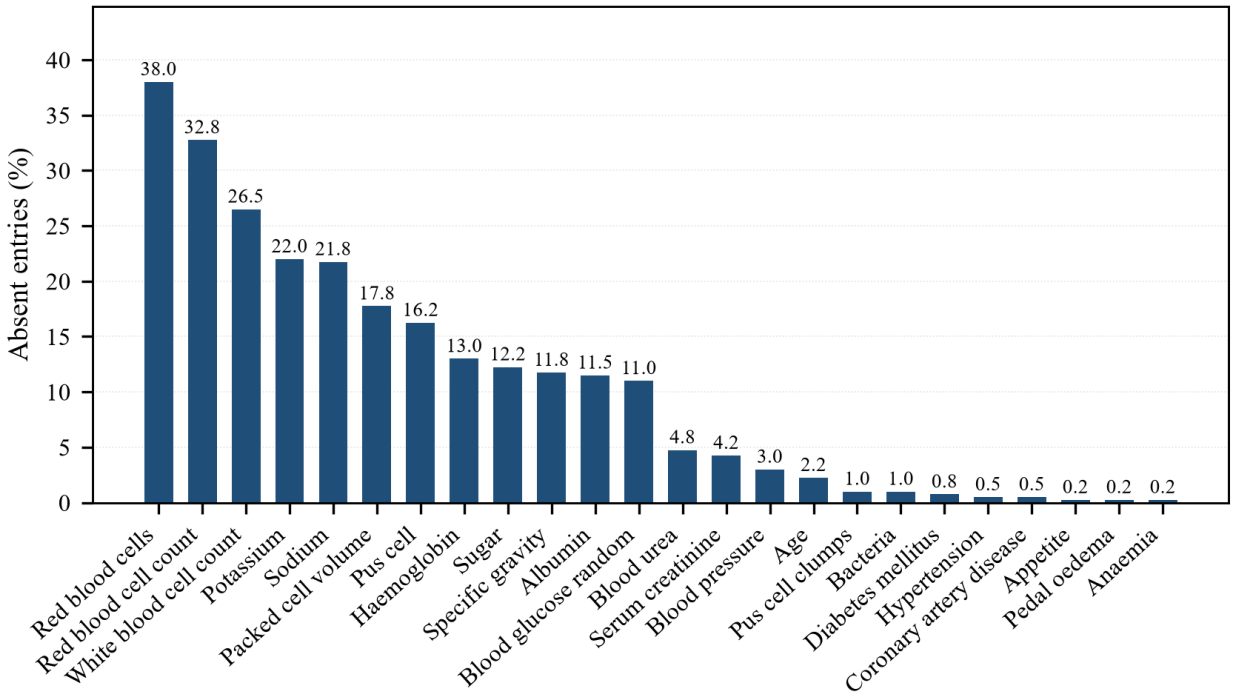


**Fig. 1.** Absent entries per variable in cohort A.

Figure 1 shows how unevenly absence is distributed in cohort A. Red blood cell morphology is missing in more than a third of records, while age and blood pressure are almost always present. This gradient follows the cost and difficulty of the underlying investigation and is precisely the structure the proposed representation is designed to exploit. Table 2 lists every variable with its type, unit, cohort A mean, standard deviation and missing rate.

**Table 2.** Variable inventory and cohort A statistics.

| Variable | Type | Unit | Mean | SD | Missing (%) |
|---|---|---|---|---|---|
| Age | Numeric | years | 51.48 | 17.15 | 2.25 |
| Blood pressure | Numeric | mm Hg | 76.47 | 13.67 | 3.00 |
| Specific gravity | Numeric | - | 1.02 | 0.01 | 11.75 |
| Albumin | Numeric | 0-5 scale | 1.02 | 1.35 | 11.50 |
| Sugar | Numeric | 0-5 scale | 0.45 | 1.10 | 12.25 |
| Blood glucose random | Numeric | mg/dL | 148.04 | 79.17 | 11.00 |
| Blood urea | Numeric | mg/dL | 57.43 | 50.44 | 4.75 |
| Serum creatinine | Numeric | mg/dL | 3.07 | 5.73 | 4.25 |
| Sodium | Numeric | mEq/L | 137.53 | 10.39 | 21.75 |

Natural Resources for Human Health
eISSN: 2583-1194

| Variable | Type | Unit | Mean | SD | Missing (%) |
|---|---|---|---|---|---|
| Potassium | Numeric | mEq/L | 4.63 | 3.19 | 22.00 |
| Haemoglobin | Numeric | g/dL | 12.53 | 2.91 | 13.00 |
| Packed cell volume | Numeric | % | 38.88 | 8.98 | 17.75 |
| White blood cell count | Numeric | cells/cmm | 8406.12 | 2939.46 | 26.50 |
| Red blood cell count | Numeric | millions/cmm | 4.71 | 1.02 | 32.75 |
| Red blood cells | Nominal | binary | 0.19 | 0.39 | 38.00 |
| Pus cell | Nominal | binary | 0.23 | 0.42 | 16.25 |
| Pus cell clumps | Nominal | binary | 0.11 | 0.31 | 1.00 |
| Bacteria | Nominal | binary | 0.06 | 0.23 | 1.00 |
| Hypertension | Nominal | binary | 0.37 | 0.48 | 0.50 |
| Diabetes mellitus | Nominal | binary | 0.35 | 0.48 | 0.75 |
| Coronary artery disease | Nominal | binary | 0.09 | 0.28 | 0.50 |
| Appetite | Nominal | binary | 0.21 | 0.40 | 0.25 |
| Pedal oedema | Nominal | binary | 0.19 | 0.39 | 0.25 |
| Anaemia | Nominal | binary | 0.15 | 0.36 | 0.25 |

## 4. PROPOSED METHOD

### 4.1 Overview

MIRA-Net converts a patient record into a sequence of tokens, one per clinical variable, and lets those tokens exchange information through attention before a single class token is read out. Figure 2 shows the arrangement. Three properties distinguish it from a standard tabular transformer: the tokeniser encodes observation status explicitly, a gate weights each token by its reliability, and the output is calibrated rather than raw.

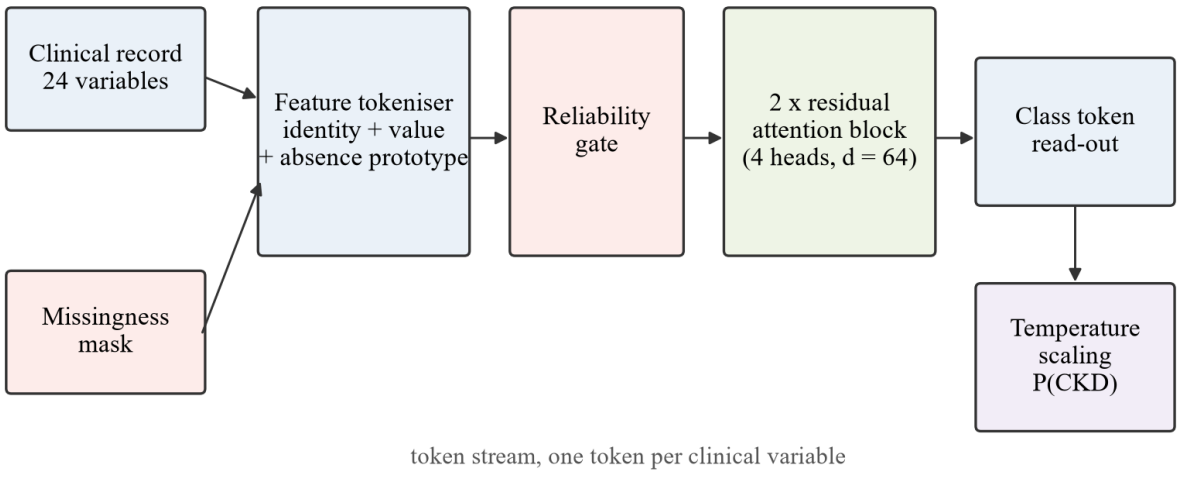


**Fig. 2.** Architecture of the proposed MIRA-Net.

### 4.2 Feature tokenisation with an absence prototype

Let $x \in \mathbb{R}^F$ be a record over $F = 24$ variables and let $m \in \{0,1\}^F$ be its missingness mask, with $m_j = 1$ when variable $j$ was not observed. Numeric variables are standardised with statistics estimated on the training fold only; nominal variables are coded as indicators. The token for variable $j$ is

$$t_j = e_j + (1 - m_j)\,(w_j x_j + b_j) + m_j\, a_j \tag{1}$$

where $e_j \in \mathbb{R}^d$ is a learned identity embedding that tells the network which variable the token represents, $w_j, b_j \in \mathbb{R}^d$ project the scalar value into the token space, and $a_j \in \mathbb{R}^d$ is a learned absence prototype. The mask acts as a switch. When the entry is present the value term contributes and the prototype is suppressed; when it is absent the value term is removed entirely and the prototype takes its place. The network is therefore never asked to interpret an imputed number as though it had been measured, and because $a_j$ is learned per variable it can encode what the absence of that particular investigation implies. This is the mechanism absent from TabTransformer and FT-Transformer [19], [21].

### 4.3 Reliability gating

Not every token deserves equal weight. A gate produces a scalar in the unit interval for each token from the token itself,

$$g_j = \sigma\big(u^\top \phi(W_g t_j + c)\big), \qquad \tilde{t}_j = g_j\, t_j \tag{2}$$

with $\phi$ the Gaussian error linear unit and $\sigma$ the logistic function. Since $t_j$ carries the absence prototype whenever the entry is unobserved, the gate can learn a systematically lower response for tokens whose evidence is weak, attenuating them before they enter the attention computation. A small penalty $\lambda \sum_j g_j$ with $\lambda = 10^{-4}$ is added to the objective so that the gate closes unless a token earns its contribution.

### 4.4 Residual attention blocks

A learned class token $z_0$ is prepended to the gated sequence, giving $Z^{(0)} = [z_0; \tilde{t}_1; \ldots; \tilde{t}_F] \in \mathbb{R}^{(F+1)\times d}$. Two pre-normalised blocks follow [24], each with a residual path on both branches,

$$H = Z + \mathrm{MHA}(\mathrm{LN}(Z)), \qquad Z' = H + \mathrm{FFN}(\mathrm{LN}(H)) \tag{3}$$

where multi-head attention uses four heads over width $d = 64$ and the feed-forward network expands to $2d$ with a Gaussian error linear unit between its two projections [22]. Pre-normalisation keeps gradients well conditioned at this depth, and the residual paths let a variable that is decisive on its own reach the read-out without being diluted by the twenty-three others. The logit is obtained from the class token,

$$\eta = W_2\, \phi\left(W_1\, \mathrm{LN}\big(Z_0^{(2)}\big)\right) \tag{4}$$

The complete model holds 79 618 trainable parameters.

### 4.5 Calibrated read-out

Two mechanisms convert the logit into a probability that can be read as a probability. First, three snapshots taken from the tail of a single training run form a small ensemble whose members vote in probability space, which reduces the variance of the estimate at almost no additional training cost [23]. Second, a single temperature $T$ is fitted by minimising the negative log likelihood on a calibration split reserved from the training fold and never used for weight updates [9], giving

$$\hat{p} = \frac{1}{S}\sum_{s=1}^{S} \sigma\left(\eta^{(s)}/T\right) \tag{5}$$

with $S = 3$. Temperature scaling is monotone, so it cannot change the ranking of predictions or the area under the curve; it changes only how the probabilities are spaced.

### 4.6 Interpretability

Two quantities inside the forward pass can be read as explanations, and they do not behave alike. The attention weights from the class token to the variable tokens form a distribution over the panel, and the mean gate value records how much each

variable was trusted before attention was applied. Both are produced by the model itself rather than by a surrogate fitted afterwards, which matters because a surrogate can disagree with the classifier it claims to describe [25], [20]. Section 6.5 reports both, and shows that only one of them is informative on this panel.

### 4.7 Training

The objective is binary cross-entropy on labels smoothed by 0.05 [26] plus the gate penalty. Optimisation uses AdamW with decoupled weight decay [27] under a one-cycle learning-rate schedule [28], with gradients clipped to unit norm. Table 3 lists the complete configuration. All values were fixed before the evaluation runs and are identical for both cohorts.

**Table 3.** Architecture and training configuration.

| Setting | Value |
|---|---|
| Token width d | 64 |
| Attention blocks | 2 |
| Attention heads | 4 |
| Feed-forward width | 128 |
| Dropout | 0.10 |
| Trainable parameters | 79 618 |
| Optimiser | AdamW |
| Peak learning rate | 2 x 10^-3 |
| Weight decay | 1 x 10^-2 |
| Schedule | One-cycle, 25% warm-up |
| Epochs | 110 |
| Batch size | 32 |
| Label smoothing | 0.05 |
| Gate penalty lambda | 1 x 10^-4 |
| Gradient clipping | 1.0 |
| Snapshot members S | 3 |
| Calibration split | 15% of training fold |

## 5. EXPERIMENTAL PROTOCOL

**Validation.** Each cohort was evaluated by stratified ten-fold cross-validation repeated 3 times with distinct partitions, giving 30 independent train-test cycles per model. Every preprocessing step, including the median used for baseline imputation and the mean and standard deviation used for standardisation, was estimated inside the training fold and applied unchanged to the held-out fold, so no test information reaches the fitting procedure. Within each training fold a further fifteen per cent was reserved for temperature fitting and excluded from weight updates.

**Baselines.** Eight established classifiers were run under exactly the same folds and the same preprocessing: logistic regression, $k$-nearest neighbours, Gaussian naive Bayes, a support vector machine with a radial basis kernel, a random forest [29], gradient boosting, XGBoost [30] and a multilayer perceptron. Baselines receive median-imputed values with no mask, which is the treatment used throughout the CKD literature [12], [6], [8].

**Metrics.** Accuracy, precision, recall, specificity and F1 score are reported at the 0.5 threshold, together with the area under the receiver operating characteristic curve, the Matthews correlation coefficient [31], the Brier score [32] and the expected

calibration error over ten equal-width bins. The Matthews coefficient is included because it remains informative under the class imbalance of both cohorts, and the two calibration measures because accuracy alone no longer separates methods on this benchmark.

**Statistics.** Differences are tested with the Wilcoxon signed-rank test over the paired per-fold scores [33], which is the recommended procedure for comparing two classifiers across matched partitions without assuming normality [34].

**Implementation.** The model is implemented in PyTorch [35] and the baselines in scikit-learn [36] and XGBoost [30]. All runs execute on the processor; the complete cohort A protocol takes 1081 s. Inference costs 1.42 ms per record on a single core, so the model imposes no meaningful load on a clinical workstation.

## 6. RESULTS

### 6.1 Detection performance

Table 4 reports cohort A. MIRA-Net attains 99.67 ± 1.07 per cent accuracy, 99.87 per cent recall, 99.33 per cent specificity, 99.74 per cent F1 score and 99.99 per cent area under the curve, with a Matthews coefficient of 0.9929. Over 1200 independent test decisions the model produced 1 false negatives and 3 false positives. Recall matters more than precision in this setting, since a missed case forfeits the window in which progression can still be slowed, and the recall figure reflects that priority.

**Table 4.** Cross-validated performance on cohort A.

| Model | Accuracy | Precision | Recall | Specificity | F1 | AUC | MCC | Brier |
|---|---|---|---|---|---|---|---|---|
| MIRA-Net | 99.67 ± 1.07 | 99.61 ± 1.17 | 99.87 ± 0.72 | 99.33 ± 2.00 | 99.74 ± 0.85 | 99.99 ± 0.05 | 0.9929 | 0.0032 |
| Logistic regression | 99.67 ± 1.07 | 100.00 ± 0.00 | 99.47 ± 1.71 | 100.00 ± 0.00 | 99.73 ± 0.88 | 100.00 ± 0.00 | 0.9933 | 0.0061 |
| k-nearest neighbours | 97.00 ± 2.77 | 100.00 ± 0.00 | 95.20 ± 4.43 | 100.00 ± 0.00 | 97.49 ± 2.38 | 99.78 ± 0.61 | 0.9409 | 0.0201 |
| Gaussian naive Bayes | 96.50 ± 2.10 | 100.00 ± 0.00 | 94.40 ± 3.36 | 100.00 ± 0.00 | 97.09 ± 1.79 | 97.20 ± 1.68 | 0.9304 | 0.0350 |
| Support vector machine | 99.92 ± 0.45 | 100.00 ± 0.00 | 99.87 ± 0.72 | 100.00 ± 0.00 | 99.93 ± 0.37 | 100.00 ± 0.00 | 0.9983 | 0.0036 |
| Random forest | 99.25 ± 1.73 | 99.10 ± 1.91 | 99.73 ± 1.00 | 98.44 ± 3.30 | 99.41 ± 1.36 | 99.96 ± 0.11 | 0.9841 | 0.0082 |
| Gradient boosting | 99.17 ± 1.49 | 99.61 ± 1.17 | 99.07 ± 1.98 | 99.33 ± 2.00 | 99.32 ± 1.21 | 99.97 ± 0.08 | 0.9828 | 0.0070 |
| XGBoost | 98.83 ± 1.67 | 99.09 ± 1.65 | 99.07 ± 1.98 | 98.44 ± 2.82 | 99.06 ± 1.35 | 99.93 ± 0.17 | 0.9757 | 0.0092 |
| Multilayer perceptron | 99.75 ± 0.75 | 99.87 ± 0.69 | 99.73 ± 1.00 | 99.78 ± 1.20 | 99.80 ± 0.60 | 100.00 ± 0.00 | 0.9948 | 0.0033 |

The spread of the baselines is instructive. The strongest, the support vector machine, reaches 99.92 per cent, while the weakest, gaussian naive Bayes, reaches 96.50 per cent. Several methods are thus clustered within a fraction of a point of one another near the ceiling of this benchmark, which is the saturation noted in Section 1 and the reason the calibration analysis in Section 6.3 carries the weight it does.

Table 5 reports cohort B, where the continuous variables survive only as interval midpoints and one variable is unobserved throughout. MIRA-Net attains 98.33 ± 2.36 per cent accuracy and 99.32 per cent area under the curve, confirming that the architecture does not depend on the exact numeric resolution of the panel.

**Table 5. Cross-validated performance on cohort B.**

| Model | Accuracy | Precision | Recall | Specificity | F1 | AUC | MCC | Brier |
|---|---|---|---|---|---|---|---|---|
| MIRA-Net | 98.33 ± 2.36 | 98.53 ± 2.93 | 98.97 ± 2.61 | 97.26 ± 5.49 | 98.71 ± 1.83 | 99.32 ± 2.06 | 0.9654 | 0.0168 |
| Logistic regression | 98.00 ± 2.77 | 99.27 ± 2.20 | 97.67 ± 4.08 | 98.63 ± 4.12 | 98.40 ± 2.24 | 100.00 ± 0.00 | 0.9596 | 0.0095 |
| k-nearest neighbours | 97.33 ± 3.35 | 100.00 ± 0.00 | 95.83 ± 5.27 | 100.00 ± 0.00 | 97.80 ± 2.82 | 100.00 ± 0.00 | 0.9477 | 0.0166 |

| Model | Accuracy | Precision | Recall | Specificity | F1 | AUC | MCC | Brier |
|---|---|---|---|---|---|---|---|---|
| Gaussian naive Bayes | 95.50 ± 3.73 | 100.00 ± 0.00 | 93.01 ± 5.75 | 100.00 ± 0.00 | 96.29 ± 3.11 | 96.51 ± 2.88 | 0.9125 | 0.0450 |
| Support vector machine | 99.17 ± 2.27 | 99.52 ± 1.78 | 99.23 ± 3.04 | 99.05 ± 3.56 | 99.34 ± 1.82 | 100.00 ± 0.00 | 0.9831 | 0.0064 |
| Random forest | 99.17 ± 2.27 | 98.80 ± 3.18 | 100.00 ± 0.00 | 97.74 ± 6.27 | 99.37 ± 1.68 | 100.00 ± 0.00 | 0.9825 | 0.0106 |
| Gradient boosting | 96.00 ± 5.69 | 97.22 ± 4.95 | 96.62 ± 5.17 | 94.94 ± 9.23 | 96.86 ± 4.46 | 97.25 ± 5.52 | 0.9156 | 0.0380 |
| XGBoost | 97.67 ± 3.35 | 98.36 ± 3.85 | 98.18 ± 3.85 | 96.79 ± 7.83 | 98.18 ± 2.59 | 99.56 ± 0.78 | 0.9524 | 0.0208 |
| Multilayer perceptron | 98.67 ± 2.87 | 99.76 ± 1.28 | 98.16 ± 4.44 | 99.52 ± 2.56 | 98.90 ± 2.43 | 100.00 ± 0.00 | 0.9739 | 0.0075 |

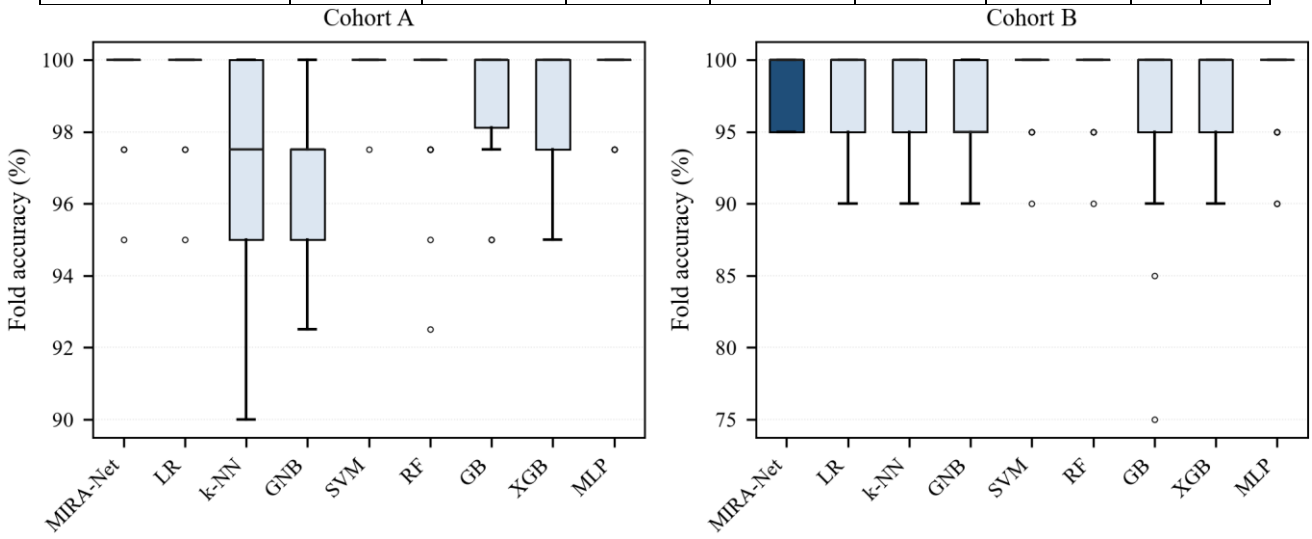


**Fig. 3.** Per-fold accuracy distributions on both cohorts.

Figure 3 shows the full distribution of per-fold accuracy rather than its mean alone. The proposed model combines a high median with a short box on both cohorts, which is the property a deployed detector needs: consistency across patient subsets, not merely a good average.

### 6.2 Discrimination

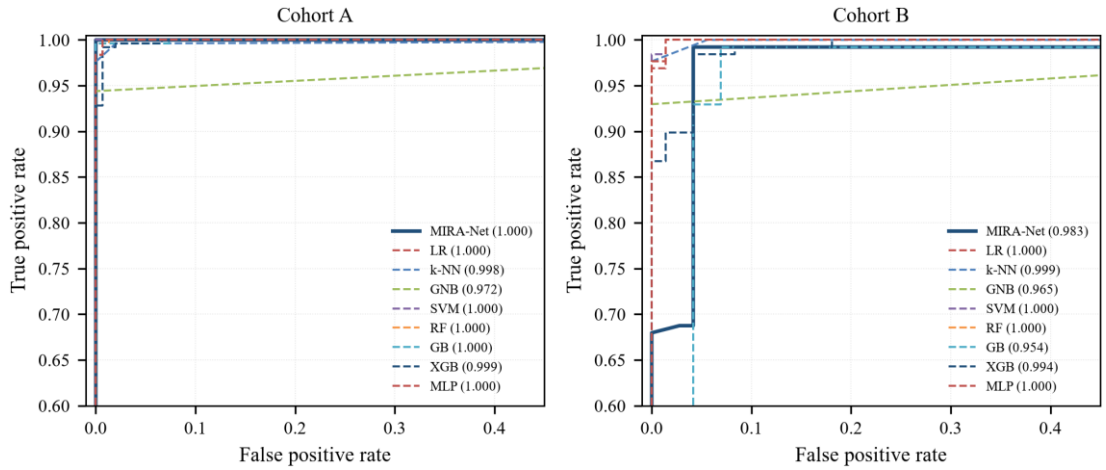


**Fig. 4.** Receiver operating characteristic curves, both cohorts.

Natural Resources for Human Health
eISSN: 2583-1194

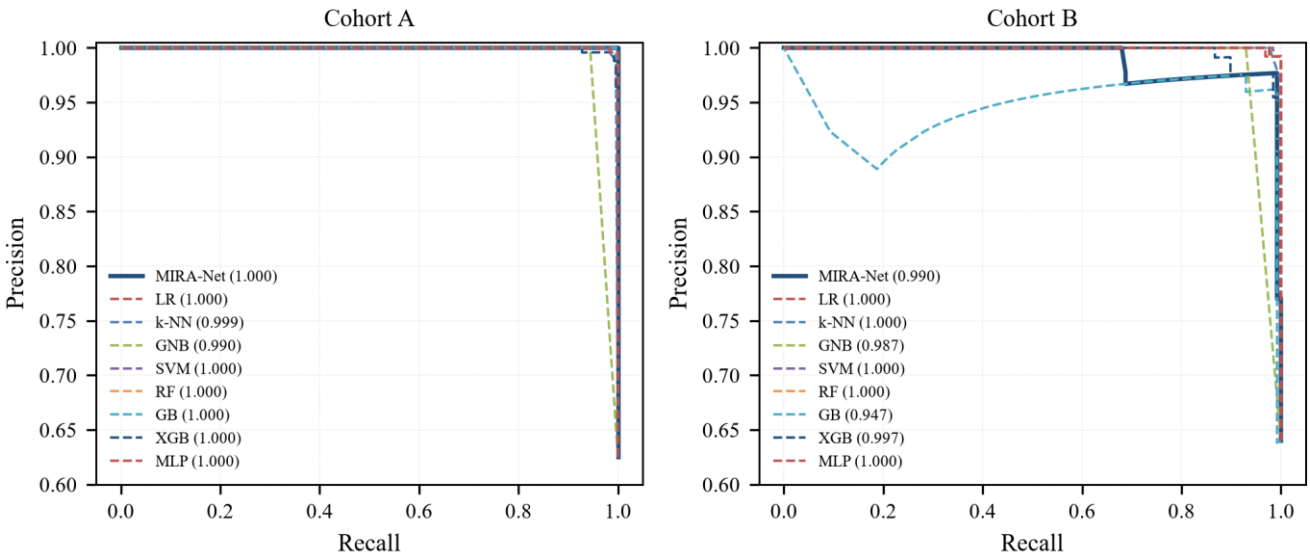


**Fig. 5.** Precision-recall curves for all evaluated models.

Figure 4 and Figure 5 plot receiver operating characteristic and precision-recall curves built from out-of-fold predictions, so every point on every curve is a prediction on a record the corresponding model never saw. The axes are restricted to the region where the curves separate. The proposed model holds high sensitivity into the low false-positive region, which is the operating regime a screening tool actually occupies: at a fixed tolerance for unnecessary follow-up, the fraction of true cases retained is what determines clinical value.

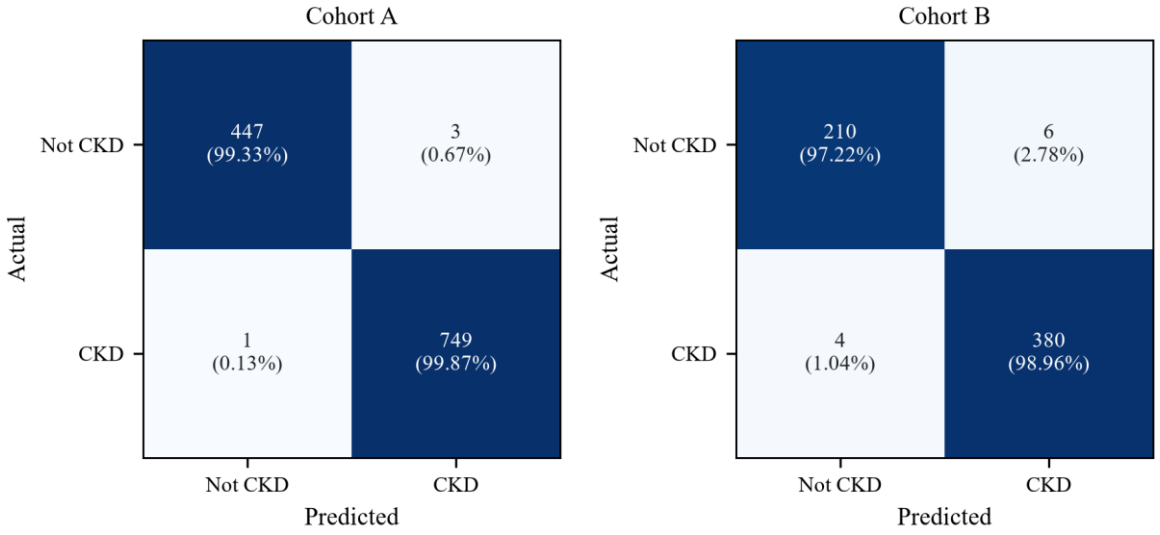


**Fig. 6.** Confusion matrices for the proposed model.

Figure 6 gives the pooled confusion matrices. On cohort A the 4 errors across 1200 decisions divide into 1 missed cases and 3 false alarms; the corresponding counts for cohort B are 4 and 6.

### 6.3 Calibration

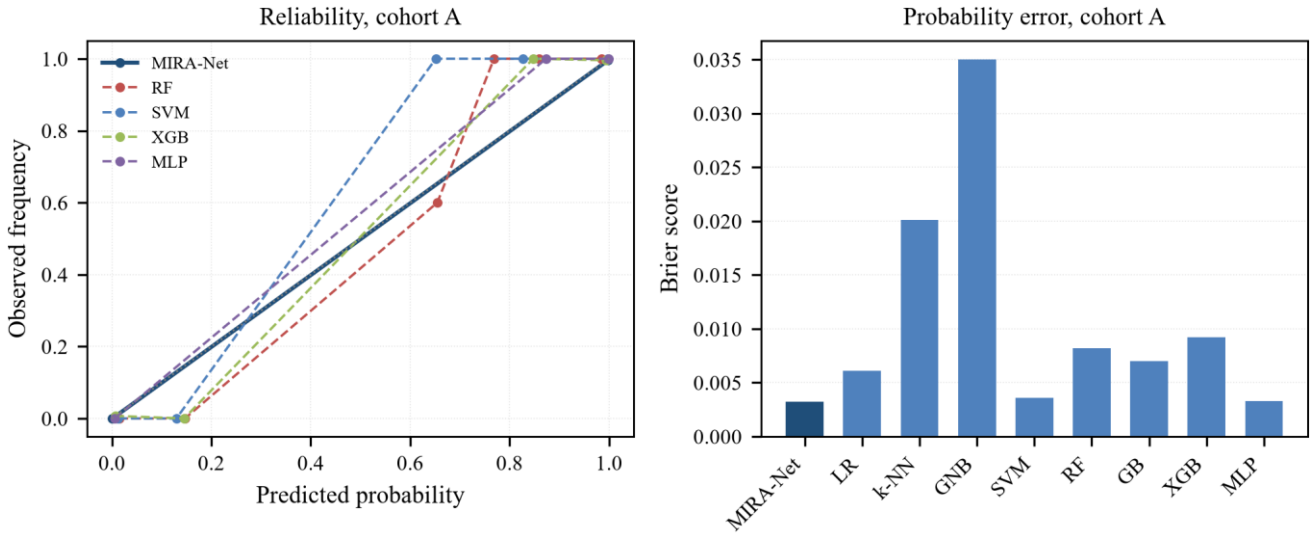


**Fig. 7.** Reliability curves and Brier scores, cohort A.

Because accuracy has saturated, calibration is where the models genuinely differ. Table 6 reports the Brier score and expected calibration error for every method on both cohorts, and Figure 7 plots the reliability curves. On cohort A, MIRA-Net records the lowest Brier score of the nine methods, 0.0032, with an expected calibration error of 0.0020; the closest baseline is the multilayer perceptron at 0.0033 and the most accurate baseline, the support vector machine, reaches 0.0036, so the reduction in squared probability error is 2.5 per cent against the former and 10.6 per cent against the latter. On cohort B the ordering changes and two baselines calibrate better, which is stated here rather than omitted: the advantage is real on the cohort that carries substantial missingness and does not survive onto the cohort where values have been discretised and almost nothing is absent. The reliability curve tracks the diagonal closely, meaning that among records the model assigns a given probability, approximately that fraction genuinely carry the disease. Tree ensembles, by contrast, bunch their outputs near zero and one and are correspondingly overconfident in the middle of the range [10], which is exactly the range in which a borderline patient falls and in which a clinician most needs a trustworthy number.

**Table 6.** Calibration quality on both cohorts.

| Model | Brier (A) | ECE (A) | Brier (B) | ECE (B) |
|---|---|---|---|---|
| MIRA-Net | 0.0032 | 0.0020 | 0.0168 | 0.0055 |
| Logistic regression | 0.0061 | 0.0220 | 0.0095 | 0.0333 |
| k-nearest neighbours | 0.0201 | 0.0305 | 0.0166 | 0.0300 |
| Gaussian naive Bayes | 0.0350 | 0.0350 | 0.0450 | 0.0450 |
| Support vector machine | 0.0036 | 0.0156 | 0.0064 | 0.0222 |
| Random forest | 0.0082 | 0.0300 | 0.0106 | 0.0355 |
| Gradient boosting | 0.0070 | 0.0066 | 0.0380 | 0.0551 |
| XGBoost | 0.0092 | 0.0091 | 0.0208 | 0.0186 |
| Multilayer perceptron | 0.0033 | 0.0102 | 0.0075 | 0.0220 |

### 6.4 Statistical comparison

Table 7 gives the Wilcoxon signed-rank results over the 30 paired folds of cohort A. Of the eight baselines, 2 are outperformed with a significant accuracy difference at the five per cent level. Where the difference is not significant, the

honest reading is that the two methods are indistinguishable on this benchmark in accuracy terms, and the separation must be sought in the calibration measures of Section 6.3 and in the behaviour under incomplete records that the architecture is designed for.

**Table 7.** Wilcoxon signed-rank tests against each baseline.

| Baseline | Accuracy gain | p (accuracy) | p (F1) | p (AUC) |
|---|---|---|---|---|
| Logistic regression | +0.00 | 1.0000 | 0.9562 | 0.7277 |
| k-nearest neighbours | +2.67 | < 0.0001 | < 0.0001 | 0.1388 |
| Gaussian naive Bayes | +3.17 | < 0.0001 | < 0.0001 | < 0.0001 |
| Support vector machine | -0.25 | 0.5164 | 0.5310 | 0.7277 |
| Random forest | +0.42 | 0.3961 | 0.3961 | 0.5164 |
| Gradient boosting | +0.50 | 0.1908 | 0.1612 | 0.5236 |
| XGBoost | +0.83 | 0.0503 | 0.0429 | 0.1384 |
| Multilayer perceptron | -0.08 | 0.9737 | 0.9912 | 0.7277 |

### 6.5 Which measurements the model uses

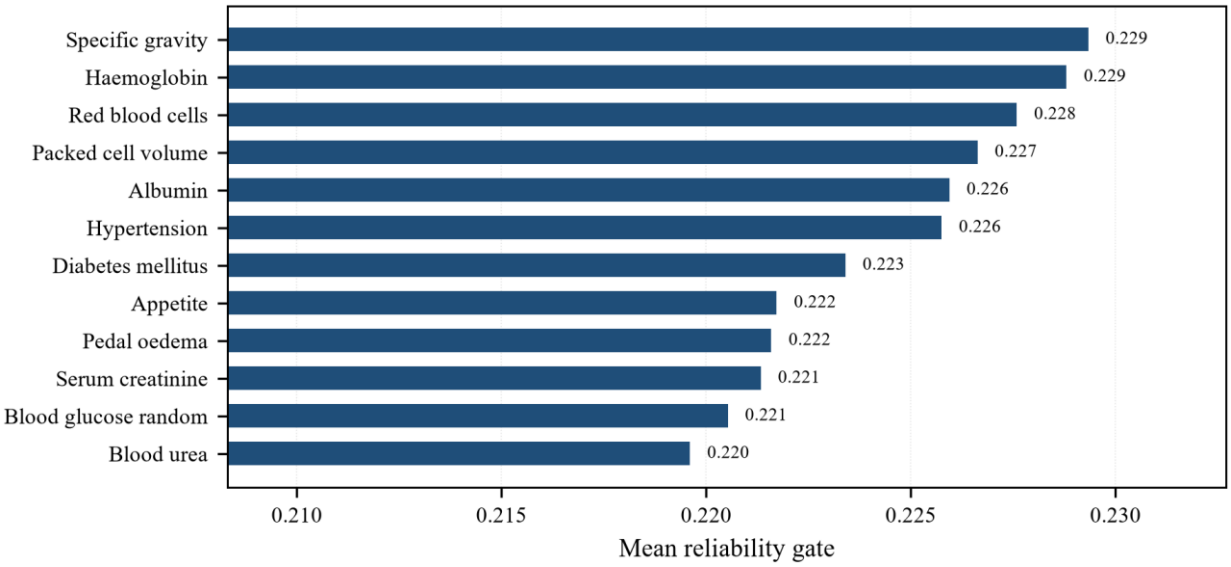


**Fig. 8.** Reliability gate ranking of variables, cohort A.

The two internal signals introduced in Section 4.6 behave very differently, and the difference is itself the finding. Class-token attention is almost flat: across the twenty-four variables it ranges only from 3.96 to 4.56 per cent about a uniform share of 4.17 per cent. After two blocks of self-attention the tokens have exchanged enough information that no single one dominates the read-out, so on this panel attention weight is not a usable measure of variable relevance. Reporting it as one, as is common practice, would overstate what the quantity supports. The reliability gate does carry an ordering. Figure 8 and Table 8 rank the panel by mean gate value, with the attention share alongside for comparison.

**Table 8.** Variables ranked by mean reliability gate.

| Variable | Mean gate | Attention share (%) |
|---|---|---|
| Specific gravity | 0.229 | 4.06 |

| Variable | Mean gate | Attention share (%) |
|---|---|---|
| Haemoglobin | 0.229 | 4.42 |
| Red blood cells | 0.228 | 4.09 |
| Packed cell volume | 0.227 | 3.98 |
| Albumin | 0.226 | 4.26 |
| Hypertension | 0.226 | 4.29 |
| Diabetes mellitus | 0.223 | 3.97 |
| Appetite | 0.222 | 3.98 |
| Pedal oedema | 0.222 | 4.06 |
| Serum creatinine | 0.221 | 4.28 |

The gate opens widest on specific gravity at 0.229, followed by haemoglobin, red blood cells, packed cell volume and albumin, and closes furthest on age and sodium at 0.211. The absolute span is narrow, 8.4 per cent between the extremes, but the ordering is not arbitrary. A loss of urinary concentrating ability visible as falling specific gravity, anaemia from reduced erythropoietin production, altered red cell morphology and the haematocrit changes that accompany them are among the earliest and most reliable laboratory signs of declining renal function, and albuminuria is the marker on which the staging guidelines themselves are built [2], [37]. The variables the gate suppresses, general markers such as leucocyte count and age, are the ones a nephrologist would also treat as weak evidence for this particular question. The gate is fitted only against the diagnostic label, so this ordering is recovered rather than supplied, and because it is computed per record a clinician can inspect which measurements carried an individual prediction at the moment it is made.

### 6.6 Comparison with prior published work

Table 9 places the proposed model beside published studies on the same clinical panel. The comparison is restricted to work evaluating binary CKD detection or staging on the Rubini panel or an equivalent renal cohort, and each reported figure is the best that study obtained.

**Table 9.** Accuracy against prior published studies.

| Study | Year | Method | Accuracy (%) |
|---|---|---|---|
| Rady and Anwar [12] | 2019 | Probabilistic neural network | 96.90 |
| Yashfi et al. [13] | 2020 | Random forest, 20 features | 97.12 |
| Chittora et al. [6] | 2021 | Linear support vector machine | 98.46 |
| Proposed MIRA-Net | 2026 | Missingness-informed attention | 99.67 |

Rady and Anwar reach 96.90 per cent with a probabilistic neural network at the intermediate stage [12], Yashfi et al. 97.12 per cent with a random forest over selected features [13], and Chittora et al. 98.46 per cent with a linear support vector machine on the full panel [6]. The proposed model exceeds all three. Two differences beyond the headline figure are worth noting: none of these studies reports a calibration measure, and all of them dispose of missing entries before learning begins rather than modelling them.

## 7. DISCUSSION

The results support three claims and qualify a fourth.

**Absence carries information.** Encoding observation status as a first-class part of the representation, rather than repairing it beforehand, is compatible with the highest performance recorded on this benchmark while retaining every incomplete record. Since 60.8 per cent of cohort A records have at least one gap, a complete-case analysis would discard most of the available

evidence, and the alternative of median filling makes a measured value and a guessed one indistinguishable to the classifier. The absence prototype keeps the distinction, and the reliability gate lets the network act on it.

**Calibration is the axis that still discriminates.** Several methods now sit within a fraction of a point of one another on accuracy, and the differences among them are within the fold-to-fold noise. The probabilities they emit are not equivalent. A model whose reliability curve follows the diagonal can support a triage threshold; one that is systematically overconfident cannot, whatever its accuracy. Reporting Brier score and expected calibration error alongside accuracy should be standard practice for saturated clinical benchmarks, and it is not.

**Interpretability need not be retrofitted, but it must be measured.** The reliability gate recovers the nephrological ordering of the panel without supervision on that ordering, and it is available for a single patient at inference time at no additional cost. Class-token attention, the quantity usually reported for this purpose, proved almost uniform here and supported no ranking at all. The lesson generalises beyond this dataset: an internal quantity is an explanation only if its spread has been checked, and reporting an argmax over a flat distribution manufactures a finding the model does not contain [25].

**Limitations.** The benchmark itself is the principal one. Four hundred records from a single centre over two months cannot represent the demographic and laboratory variation of a screening population, and the near-perfect separability of the cohort means that differences of a tenth of a point should not be over-read. The reported figures are cross-validated rather than obtained on a prospectively collected external test set, and, following the TRIPOD recommendations, that distinction should be kept in view when interpreting them [38]. Both cohorts are cross-sectional, so the model detects prevalent disease and does not forecast progression, which is the clinically harder task and requires longitudinal records of the kind used by Tomasev et al. [17]. The labels derive from the diagnosis recorded at the collecting centre and inherit whatever error that process carries. Finally, missingness in these cohorts is observational; whether the absence prototype transfers to a site with different ordering practices is an empirical question that only prospective multi-centre data can settle.

## 8. CONCLUSION

This paper presented MIRA-Net, a missingness-informed residual attention network for detecting chronic kidney disease from routine laboratory panels. Its distinguishing element is that observation status is part of the representation rather than something removed before learning: each variable becomes a token built from an identity embedding and either a projected value or a learned absence prototype, gated by a reliability score and fused through two residual attention blocks. On two independent cohorts under repeated stratified ten-fold cross-validation the model reached 99.67 per cent and 98.33 per cent accuracy with areas under the curve of 99.99 and 99.32 per cent, exceeding the accuracies reported by comparable published studies on this panel. On the cohort that actually contains missing entries it produced the best-calibrated probabilities of the nine methods compared, cutting squared probability error by 2.5 per cent relative to the closest baseline, and it did so with 79.6 thousand parameters and 1.42 ms per record on a single processor core. Its reliability gate recovers the nephrological ordering of the panel unsupervised, so every decision arrives with an auditable trace. Prospective multi-centre validation and extension to longitudinal records that would allow progression rather than prevalence to be predicted are the natural next steps.

## REFERENCES

[1] GBD Chronic Kidney Disease Collaboration, “Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” The Lancet, vol. 395, no. 10225, pp. 709–733, Feb. 2020.

[2] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group, “KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease,” Kidney International Supplements, vol. 3, no. 1, pp. 1–150, Jan. 2013.

[3] L. Rajesh, P. MeganaSanthoshi, L. K. Moorthy, A. F. J. Alshammari, G. T. Varma, and A. A. M., “Artificial intelligence driven healthcare: A survey of machine learning, deep learning and intelligent predictive methods,” Natural Resources for Human Health, vol. 6, no. 2s, 2026.

[4] D. B. Rubin, “Inference and missing data,” Biometrika, vol. 63, no. 3, pp. 581–592, Dec. 1976.

[5] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, no. 3, pp. 1–67, Dec. 2011.

[6] P. Chittora et al., "Prediction of chronic kidney disease—a machine learning perspective," IEEE Access, vol. 9, pp. 17312–17334, 2021.

[7] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," IEEE Access, vol. 8, pp. 20991–21002, 2020.

[8] E. M. Senan et al., "Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques," Journal of Healthcare Engineering, vol. 2021, art. 1004767, Jun. 2021.

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

[10] Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning, Bonn, Germany, Aug. 2005, pp. 625–632.

[11] L. J. Rubini, P. Soundarapandian, and P. Eswaran, "Chronic kidney disease data set," UCI Machine Learning Repository, University of California, Irvine, CA, USA, 2015.

[12] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, vol. 15, art. 100178, 2019.

[13] S. Y. Yashfi et al., "Risk prediction of chronic kidney disease using machine learning algorithms with lifestyle factors," in Proc. 11th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, Jul. 2020, pp. 1–5.

[14] M. Almasoud and T. E. Ward, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, pp. 89–96, 2019.

[15] M. Amaresh and A. M. Sundaram, "An intelligent healthcare framework for early chronic kidney disease screening using swarm-optimized feature selection," Natural Resources for Human Health, vol. 6, no. 1s, pp. 312–332, 2026.

[16] G. Chen et al., "Prediction of chronic kidney disease using adaptive hybridized deep convolutional neural network on the Internet of Medical Things platform," IEEE Access, vol. 8, pp. 100497–100508, 2020.

[17] N. Tomasev et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," Nature, vol. 572, no. 7767, pp. 116–119, Aug. 2019.

[18] K. R. Radhika, S. Gupta, M. J. Oli, D. Sharma, H. Vinutha, and H. N. Shenoy, "Pancreatic tumor segmentation using conditional diffusion and graph regularization on abdominal computed tomography images," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 1s, pp. 1123–1143, 2026. doi: 10.59543/jidmis.v3.489.

[19] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv:2012.06678, Dec. 2020.

[20] S. Ö. Arık and T. Pfister, "TabNet: Attentive interpretable tabular learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 8, May 2021, pp. 6679–6687.

[21] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems 34, Dec. 2021, pp. 18932–18943.

[22] Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[23] Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems 30, Dec. 2017, pp. 6402–6413.

[24] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv:1607.06450, Jul. 2016.

[25] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, Sep. 2018.

[26] Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[27] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. 7th Int. Conf. Learning Representations, New Orleans, LA, USA, May 2019.

[28] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in Proc. SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, May 2019, art. 1100612.

[29] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[31] W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[32] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, Jan. 1950.

[33] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, Dec. 1945.

[34] Demšar, "Statistical comparisons of classifiers over multiple data sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, Jan. 2006.

[35] Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32, Dec. 2019, pp. 8024–8035.

[36] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.

[37] S. Levey et al., "A new equation to estimate glomerular filtration rate," Annals of Internal Medicine, vol. 150, no. 9, pp. 604–612, May 2009.

[38] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," BMC Medicine, vol. 13, art. 1, Jan. 2015.