STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–11.
Published online in International Academic Press (www.IAPress.org)

An Optimized Deep Learning Framework for Pension Fund Risk Prediction: A Rolling-Origin Evaluation on Indonesian Defined-Benefit Pension Fund Data

Wahyuni Ekasasmita [1,*], Nur Rahmi [2], Rahmat H.S [3], Ahmad Fajri [1],

[1]*Department of Actuarial Science, Institut Teknologi Bacharuddin Jusuf Habibie, Indonesia*
[2]*Department of Mathematics, Institut Teknologi Bacharuddin Jusuf Habibie, Indonesia*
[3]*Department of Statistics, Universitas Negeri Makassar, Indonesia*

Abstract
Deep learning is often proposed for early warning of pension fund stress, but rarely tested against simple benchmarks on short monthly series. We build OPFR-Net, a multi-output neural network that jointly predicts next-month funding ratio, solvency risk and liquidity risk, with architecture chosen by nested random search. Using a documented monthly dataset of Indonesian defined-benefit funds from OJK statistics (2017-2026), we evaluate all models by expanding-window one-step-ahead forecasting over 57 out-of-sample months, with five seeds and Diebold–Mariano tests. The result is negative: persistence (RMSE 0.74 percentage points) and AR(1) (0.73) are most accurate, the tuned OPFR-Net reaches 0.87 ± 0.05, and the standard network is significantly worse than persistence. Macroeconomic variables and the binary risk heads add no value. A positive control confirms the pipeline detects a planted signal. OPFR-Net is offered as an open, leakage-free evaluation framework, not as a more accurate predictor.

Keywords Pension fund, Deep learning, Financial risk management, Hyperparameter optimization, Actuarial risk, Forecast evaluation

AMS 2010 subject classifications 62P05, 62P20, 68T07, 90C90, 91G70

DOI: 10.19139/soic-2310-5070-4563

1. Introduction

Defined-benefit (DB) pension funds promise benefits that are fixed by formula, so the fund, not the member, carries the risk that assets fall short of the actuarial value of liabilities. Supervisors and trustees therefore monitor the funding ratio (FR), the ratio of net assets to the actuarial present value of benefits, together with liquidity indicators such as the number of months of benefit payments covered by cash-like assets. Funding stress builds up gradually and is driven by investment returns, contribution and benefit flows, asset-allocation shifts and interest-rate movements, which interact over time. This is the usual motivation for data-driven early-warning models, and for deep neural networks in particular [5, 6, 7].

The pension literature is dominated by stochastic-control and optimisation models of funding and investment [1, 8, 9, 10, 11]. These models are valuable but need distributional assumptions, and they are not designed to be validated month by month against supervisory data. Empirical machine-learning studies, on the other hand, frequently evaluate on a random split of a short, autocorrelated series and compare a neural network only with weak benchmarks. That practice is known to overstate the value of machine-learning forecasters [12, 13], and on

*Correspondence to: Wahyuni Ekasasmita (Email: wahyuni.ekasasmita@ith.ac.id). Department of Actuarial Science, Institut Teknologi Bacharuddin Jusuf Habibie. Parepare, Indonesia (91121).

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

medium-sized tabular data neural networks are not generally superior to tree ensembles [14]. An earlier version of this manuscript suffered from these weaknesses: it reported results on 75 rows of data that could not be traced to a public source, used a random split, and claimed feature and hyper-parameter optimisation that its code did not perform. We rebuilt the study from public regulatory statistics and a leakage-free protocol, and the main contribution of this version is to show, with a transparent and reproducible workflow, what deep learning does and does not deliver in this setting.

(RQ1) Can a tuned multi-output neural network (OPFR-Net) forecast next-month funding ratio more accurately than persistence, AR(1), ridge regression and tree ensembles? (RQ2) Do its binary solvency and liquidity heads beat the trivial rule "next month's label equals this month's label"? (RQ3) Do macroeconomic variables (policy rate, inflation) improve forecasts? (RQ4) Are the explainability and scenario outputs of the network stable enough to support supervisory interpretation? (RQ5) Would the pipeline detect a genuine signal if one existed?.

We construct and document a monthly dataset of Indonesian DB (PPMP) pension funds, January 2017 - March 2026, directly from OJK statistical files and from Bank Indonesia and national statistics for macroeconomic variables. Cleaning rules, such as de-accumulating year-to-date statements and handling a restated benefit figure, are explicit and scripted. OPFR-Net is a multi-output network with a regression head for the funding ratio change and sigmoid heads for solvency and liquidity risk. A nested random search selects architecture, regularisation and loss weight with an objective that is independent of the searched loss weight, so the search cannot trivially favour one task. The model is implemented in NumPy and verified by a numerical gradient check. Expanding-window one-step-ahead forecasting, chronological validation, benchmark models that include persistence and AR(1), five seeds, Diebold-Mariano tests with the Harvey-Leybourne-Newbold correction, an ablation of each design element, and a positive-control experiment. On these data no learned model beats persistence; macroeconomic variables, the classification heads and the importance and sensitivity outputs are small and do not indicate predictive information. We state when such conclusions could change (more funds, longer horizons, richer liability data). Open data and code that reproduce every table and figure of this paper.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the data. Section 4 presents the OPFR-Net framework, the hyper-parameter search, the evaluation protocol and the statistical tests. Section 5 reports the results. Section 6 discusses implications and limitations, and Section 7 concludes.

2. Related work

A large body of work formulates pension funding and investment as stochastic optimisation, for example mean–variance funding of defined-contribution plans with administrative costs and taxation [1], target-based mean–variance control [10], robust optimisation of fund management [8], optimal funding of DB plans with stochastic mortality and interest rates [9], and time-consistent investment and contribution decisions for DB plans [11]. These papers derive optimal policies given a model of the market. We address a complementary, empirical question: how well can the observed funding position be forecast one month ahead from published supervisory data?

Deep learning has been applied to mortality, reserving, pricing and other actuarial tasks [5, 6], and its statistical foundations are covered in [7]. Recent papers in this journal compare machine-learning models with classical ones in financial and insurance settings, for example ARIMA, support vector regression and LSTM for international stock-market indices [2] and machine-learning models for auto-insurance claims [3]. Forecasting of nonstationary time series, with a discrete Hilbert transform approach, has also been studied in this journal by some of the present authors [4]; that work addresses nonstationarity in the signal itself, whereas here the focus is the comparison of learned models with simple benchmarks. We follow this line of comparative evaluation but apply it to pension funding indicators, where the series are short and strongly persistent.

Makridakis et al. [12] found that popular machine-learning methods were less accurate than simple statistical methods on 1045 monthly M3 series, and Hewamalage et al. [13] document common pitfalls, such as weak benchmarks, data leakage and unsuitable data splits. Grinsztajn et al. [14] show that tree-based models remain strong on medium-sized tabular data even after tuning. Rolling-origin evaluation [15] and tests of equal predictive accuracy [16, 17] are the standard tools to guard against these problems and are the basis of our protocol.

Table 1. Variables. Features are observed at month t; targets refer to month $t+1$.

Variable	Definition
Funding ratio	$\mathrm{FR}_t = A_t/L_t$ (net assets / actuarial present value of benefits)
Monthly return, volatility	investment income$_t$ / total assets$_{t-1}$ (%); rolling 6-month standard deviation
Contribution, benefit rate	monthly contributions, benefit payments as % of previous-month assets
Liquidity cover	(cash + on-call deposits) / mean benefit payments of last 3 months (months)
Portfolio mix	shares of government bonds, listed equities and time deposits in investments (%)
BI rate, inflation	policy rate and CPI inflation y/y (from 2020-01 only)
Solvency risk	1 if $\mathrm{FR} < 1.00$
Liquidity risk	1 if liquidity cover < 1.5 months

Random search is an efficient baseline for neural-network hyper-parameters [18]. For interpretation, permutation importance [22, 25] and Shapley-value attributions [26] are widely used. We use permutation importance because it is what our implementation computes, and report its sampling and seed variability explicitly; we do not claim SHAP values.

3. Data

3.1. Sources and scope

The data cover conventional pension funds of the Defined Benefit type (*Program Pensiun Manfaat Pasti*, PPMP) at monthly frequency from January 2017 to March 2026 (111 months). They are aggregated at industry level. Balance-sheet items (net assets, actuarial present value of benefits), investment income, contributions, benefit payments and the investment portfolio come from the OJK pension-fund statistics (snapshot of 28 May 2026) [27]; the policy rate (BI 7-Day Reverse Repo Rate) comes from Bank Indonesia [28] and consumer-price inflation from Statistics Indonesia [29]. The macroeconomic series are available in our files only from January 2020.

3.2. Variable definitions

Let A_t be net assets (*Aset Neto*) and L_t the actuarial present value of PPMP benefits. The funding ratio is $\mathrm{FR}_t = A_t/L_t$. Table 1 lists all variables. OJK income and cash-flow statements are cumulative within a calendar year; we de-accumulate them each January. One year-to-date benefit figure (November 2021) is restated downwards by OJK; the resulting two-month total is spread evenly over November and December 2021. Monthly return is investment income divided by the previous month's total assets.

3.3. Targets and descriptive statistics

The regression target is FR_{t+1} (modelled as the change $\Delta\mathrm{FR}_{t+1}$). The binary targets use two thresholds that are explicit modelling assumptions, not regulatory facts nor fitted values: $\mathrm{FR} < 1$ for solvency risk and liquidity cover below 1.5 months for liquidity risk. Because the thresholds are fixed in advance, the labels are never derived from sample quantiles. Table 2 gives descriptive statistics. The funding ratio varies between 0.90 and 1.05 with mean 0.98; 80.2% of months have $\mathrm{FR} < 1$ and 32.1% have low liquidity cover (Figure 1). Solvency risk is therefore highly imbalanced, and in the test period only 5 of 57 months have $\mathrm{FR} \geq 1$.

Table 2. Descriptive statistics (monthly, 2017-01 to 2026-03; the first rolling-window values and the macro series are shorter).

Variable	n	Mean	SD	Min	Max
Funding ratio (FR)	111	0.980	0.027	0.897	1.049
Shortfall ratio	111	0.024	0.022	0.000	0.103
Monthly investment return (					
6-month return volatility	106	0.100	0.068	0.013	0.312
Contribution rate (					
Benefit rate (					
Liquidity cover (months)	109	1.756	0.569	0.958	5.826
Government-bond share (					
Equity share (					
Time-deposit share (					
BI 7-day RR (					
Inflation y/y (					
Net assets (Rp trillion)	111	168.838	14.499	147.614	195.438

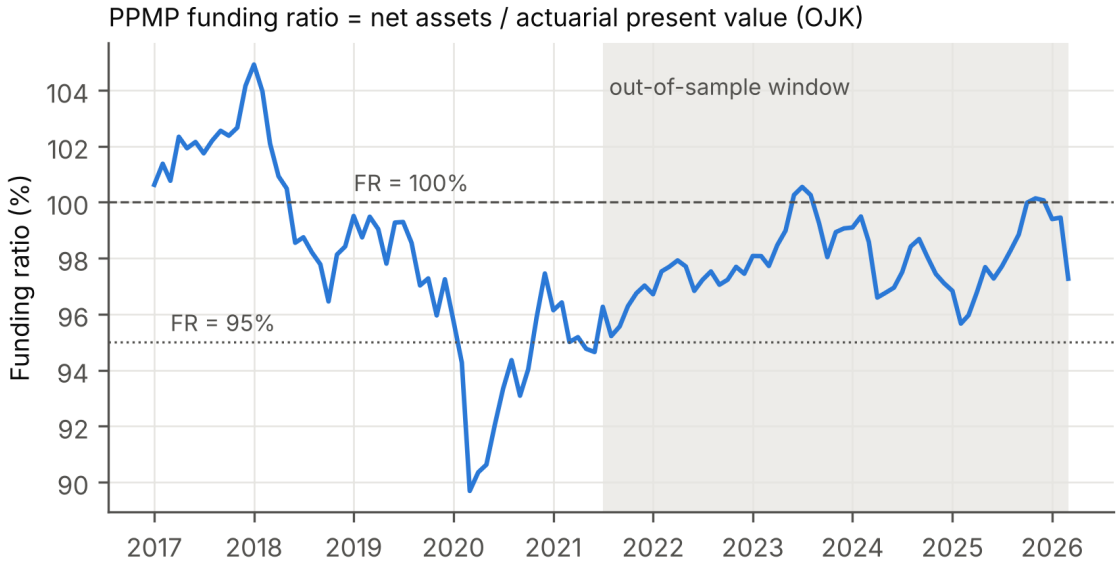


Figure 1. Funding ratio of Indonesian PPMP funds (OJK data), with the $\mathrm{FR} = 100\%$ threshold used for the solvency label.

4. The OPFR-Net framework and evaluation protocol

4.1. Problem setting

Let $\mathbf{x}_t \in \mathbb{R}^p$ be the features at month t (nine OJK-derived variables; two macroeconomic variables are added in one ablation). The targets at $t+1$ are the funding ratio y^{FR}_{t+1} and two binary indicators $y^{\mathrm{sol}}_{t+1}, y^{\mathrm{liq}}_{t+1}$. The forecasting problem is

$$\widehat{\mathrm{FR}}_{t+1} = \mathrm{FR}_t + g_\theta(\mathbf{x}_t), \qquad \hat{p}^{(j)}_{t+1} = \sigma\big(h^{(j)}_\theta(\mathbf{x}_t)\big),\ j \in \{\mathrm{sol}, \mathrm{liq}\}, \tag{1}$$

so the network learns the change in funding ratio and is anchored on persistence ($g_\theta \equiv 0$ gives the persistence forecast).

4.2. Architecture and loss

OPFR-Net is a feed-forward network with L hidden layers, $\mathbf{h}^{(0)} = \mathbf{x}$ (standardised with training-sample moments) and

$$\mathbf{h}^{(l)} = \mathrm{drop}\big(\varphi(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})\big), \quad l = 1, \dots, L, \tag{2}$$

with $\varphi \in \{\mathrm{ReLU}, \tanh, \mathrm{GELU}\}$ [21] and dropout [20]. A linear regression head outputs the standardised change g_θ and two sigmoid heads output the risk probabilities. Training minimises

$$\mathcal{L}(\theta) = \alpha \frac{1}{n}\sum_i \big(\Delta\mathrm{FR}_i - g_\theta(\mathbf{x}_i)\big)^2 - (1-\alpha)\frac{1}{2n}\sum_i\sum_j \Big[y_i^{(j)} \log \hat{p}_i^{(j)} + (1-y_i^{(j)})\log(1-\hat{p}_i^{(j)})\Big] + \gamma\|\mathbf{W}\|_2^2, \tag{3}$$

with Adam [19], mini-batches and early stopping on a chronological validation block. The regression target is standardised with training moments. The network is implemented in NumPy; its analytic gradients agree with finite differences to a maximum relative error of 9.5×10^{-9}.

4.3. Hyper-parameter optimisation

Let $\lambda = (L, \text{widths}, \varphi, \text{dropout}, \text{learning rate}, \gamma, \alpha, \text{batch size})$ with $L \in \{1, 2, 3\}$, 8–64 units, dropout in $[0, 0.3]$, learning rate in $[10^{-3}, 10^{-2}]$, $\gamma \in [10^{-5}, 10^{-2}]$, $\alpha \in [0.5, 0.9]$ and batch size in $\{8, 16, 32\}$. We use random search [18] with 20 trials. Each configuration is trained on the training rows and scored on the last 12 rows (chronological validation) by

$$J(\lambda) = \tfrac{1}{2}\,\mathrm{MSE}_{\mathrm{val}}(\Delta\mathrm{FR}) + \tfrac{1}{2}\,\mathrm{BCE}_{\mathrm{val}} + \kappa\, P(\lambda), \qquad \kappa = 2 \times 10^{-6}, \tag{4}$$

where $P(\lambda)$ is the number of trainable parameters. The weights $\frac{1}{2}, \frac{1}{2}$ in (4) are fixed and do not involve the searched α; the selection criterion therefore cannot be reduced by simply shifting weight between tasks. The search is repeated every 12 forecast origins, using only data available at that time. The selected configurations (one per block) are stored with the results; they change from block to block (from one hidden layer of 8 units to three layers of up to 64 units), a sign that the validation signal is weak.

4.4. Evaluation protocol

Models are evaluated by expanding-window one-step-ahead forecasting. For each origin k the model is trained on rows before k only, the last 12 of which serve as validation, then refitted on all earlier rows and used to predict month $k+1$. The first origin has 48 training rows, giving 57 out-of-sample months (July 2021 – March 2026). In the macro ablation the window starts in January 2020 (36 initial rows, 38 out-of-sample months). Neural networks are trained with five seeds; tables report means and standard deviations over seeds, and tests use the seed-averaged forecast.

Persistence ($\widehat{\mathrm{FR}}_{t+1} = \mathrm{FR}_t$), expanding mean, AR(1) on FR, ridge regression [24], random forest [22] and gradient boosting [23], with hyper-parameters tuned in the same blocks. For classification: label persistence, majority class and logistic regression. Neural variants are: a standard DNN (128-64-32, ReLU), the original OPFR-Net configuration (256-80-16, GELU), a small fixed OPFR-Net (32-16, GELU), the same without regularisation, a regression-head-only variant, and the tuned OPFR-Net.

Forecast accuracy is measured by MAE and RMSE in percentage points (pp) of the funding ratio, and by the MSE skill relative to persistence, $1 - \mathrm{MSE}_m/\mathrm{MSE}_{\mathrm{pers}}$ (negative: worse than persistence). Equal predictive accuracy is tested with the Diebold–Mariano test [16] for horizon $h = 1$ on squared-error differences with the Harvey–Leybourne–Newbold small-sample correction [17], against a Student t_{n-1} reference. Classifiers are scored by accuracy, balanced accuracy, AUC and Brier score.

Permutation importance [22, 25] is the increase in out-of-sample RMSE when a variable (or a group of related variables) is permuted across the test months, averaged over 30 permutations and reported separately for each of five seeds. Scenario sensitivity shifts one input at a time (for example investment return -10%) and records the change in the forecast funding ratio. Both are compared to seed-to-seed variability and to the forecast error, so that effects within noise are not interpreted.

Table 3. One-step-ahead funding-ratio forecasts, July 2021 – March 2026 (57 months). ± is the standard deviation over five seeds. MSE skill is relative to persistence. DM p: Diebold–Mariano test against persistence.

Model	MAE (pp)	RMSE (pp)	MSE skill	DM p
Persistence	0.58	0.74	0.00	–
Expanding mean	1.03	1.28	−1.98	<0.001
AR(1)	0.59	0.73	0.03	0.411
Ridge	0.59	0.76	−0.06	0.122
Random forest	0.58	0.75	−0.04	0.626
Gradient boosting	0.62	0.80	−0.17	0.062
Standard DNN	0.69 ± 0.02	0.88 ± 0.03	-0.43 ± 0.11	0.010
OPFR-Net (paper config)	0.66 ± 0.03	0.84 ± 0.03	-0.28 ± 0.08	0.039
OPFR-Net small fixed	0.70 ± 0.04	0.89 ± 0.07	-0.45 ± 0.24	0.272
- no regularisation	0.72 ± 0.05	0.92 ± 0.08	-0.58 ± 0.29	0.222
- regression head only	0.70 ± 0.04	0.90 ± 0.07	-0.48 ± 0.24	0.386
OPFR-Net tuned	0.66 ± 0.02	0.87 ± 0.05	-0.41 ± 0.17	0.186

To check that the pipeline can detect a real signal, we replace the target by $\mathrm{FR}_{t+1} = \mathrm{FR}_t + 0.010\, z(\text{return}_t) + 0.008 \tanh(2z(\text{benefit rate}_t)) + \varepsilon_t$, $\varepsilon_t \sim N(0, 0.003^2)$, in the same features and run the identical protocol.

5. Results

5.1. *Funding-ratio forecasts (RQ1)*

Table 3 and Figures 2–3 summarise the 57 out-of-sample months. Persistence (RMSE 0.74 pp) and AR(1) (0.73 pp) are the most accurate; ridge (0.76), random forest (0.75) and gradient boosting (0.80) are statistically indistinguishable from persistence (DM $p = 0.122$, 0.626, 0.062). The tuned OPFR-Net has RMSE 0.87 ± 0.05 pp (MSE skill -0.41, $p = 0.186$). The standard DNN (0.88 ± 0.03) and the original OPFR-Net configuration (0.84 ± 0.03) are significantly worse than persistence ($p = 0.010$ and 0.039). Removing regularisation or the classification heads does not help (0.92 and 0.90), and the tuned network is not better than the fixed small network (0.87 vs. 0.89, the difference is within seed variability). We therefore find no evidence that optimisation or the multi-output design improves funding-ratio forecasts on these data. With 57 test months and several comparisons, we do not interpret differences of a few hundredths of a percentage point.

The mean forecast error is small for all leading models (-0.05 pp for persistence, -0.12 pp for AR(1), -0.004 pp for the tuned OPFR-Net), so there is no systematic under-prediction. MAE and RMSE rank the models consistently (the neural networks have a higher RMSE and MAE than persistence); the largest single-month error is 2.2 pp for both persistence and the tuned network, so the network does not reduce occasional large misses. Models are tuned under the same expanding-window protocol, and preprocessing (standardisation) is fitted on training rows only. Rows with missing values (the first months of rolling windows, and the macro series before 2020) are excluded rather than imputed.

5.2. *Macroeconomic variables (RQ3)*

Table 4 adds the BI rate and inflation on the shorter 2020–2026 window. Persistence has RMSE 0.79 pp in both settings. The tuned OPFR-Net moves from 0.94 ± 0.06 to 0.91 ± 0.05, a change smaller than the seed standard deviation and still worse than persistence; for the small fixed network the RMSE rises from 0.94 to 1.06. The macroeconomic variables therefore do not improve forecasts in our sample.

5.3. *Risk classification (RQ2)*

Table 5 reports the binary heads and benchmarks. For solvency risk, persistence of the label has balanced accuracy 0.78, whereas the neural networks are at or below chance (balanced accuracy 0.44–0.54, AUC 0.29–0.49). The test

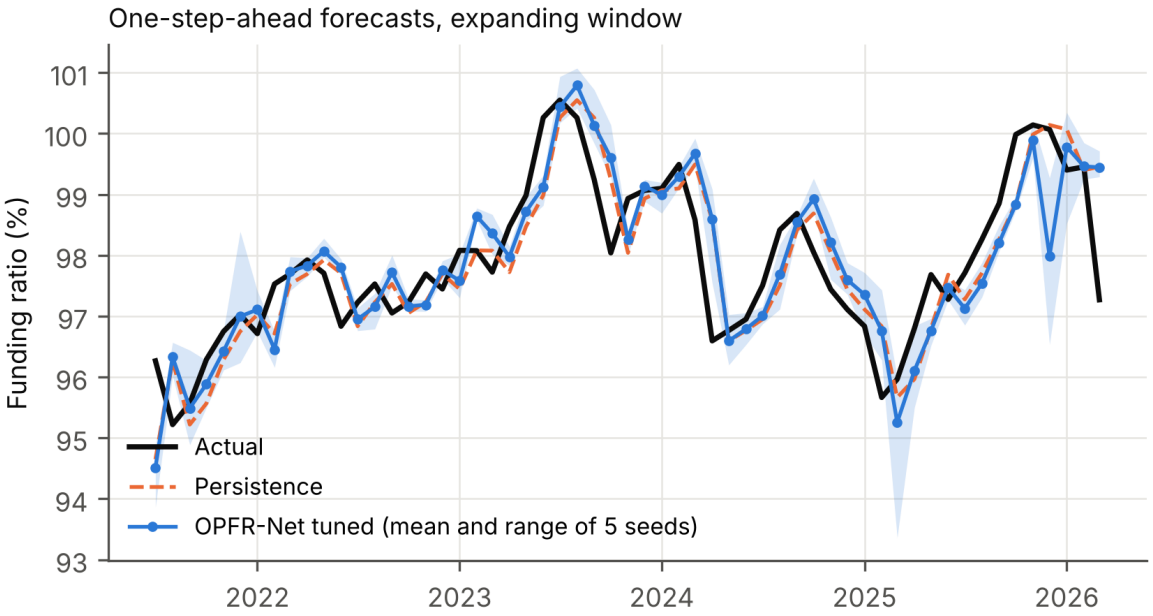


Figure 2. Actual and forecast funding ratio in the out-of-sample period.

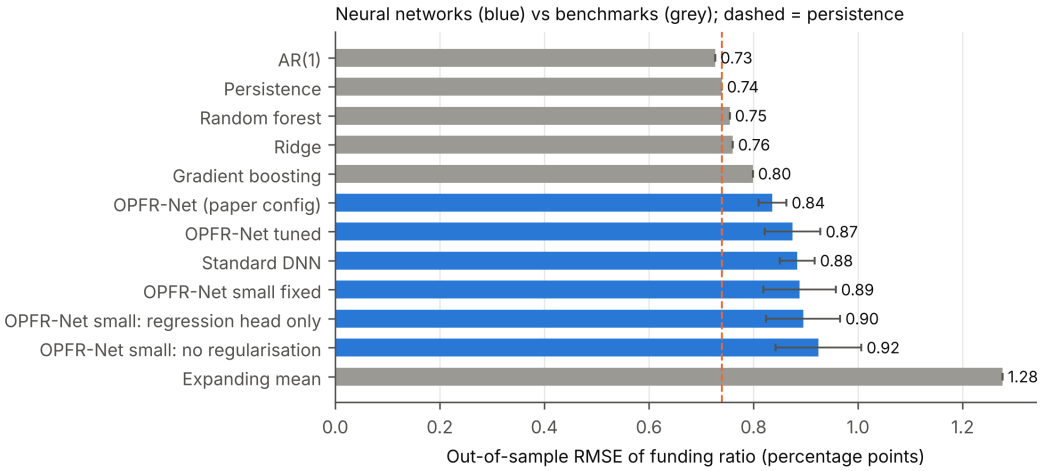


Figure 3. Out-of-sample RMSE by model (percentage points of funding ratio); error bars show the standard deviation over seeds for neural networks.

set contains only five months with FR ≥ 1, so any AUC is unreliable for this target, and the majority-class accuracy (0.91) is already high. For liquidity risk, label persistence (balanced accuracy 0.66, AUC 0.66) and random forest (0.60, 0.66) are best, while the neural networks are close to chance (AUC 0.49–0.55). The classification heads therefore add no value over label persistence.

5.4. Explainability and sensitivity (RQ4)

Table 6 and Figure 4 show permutation importances. The largest effect is for the portfolio-mix group (+0.077 pp of RMSE, about 9% of the model's error), driven by the government-bond share (+0.075 pp); both are positive in

Table 4. Macro ablation (window starting 2020-01; 38 out-of-sample months).

Feature set / model	RMSE (pp)	MSE skill	DM p
OJK only			
Persistence	0.79	0.00	–
AR(1)	0.79	−0.00	0.988
Ridge	0.82	−0.08	0.187
Random forest	0.86	−0.18	0.108
OPFR-Net small fixed	0.94 ± 0.04	-0.43 ± 0.12	0.150
OPFR-Net tuned	0.94 ± 0.06	-0.41 ± 0.18	0.065
OJK + BI rate + inflation			
Persistence	0.79	0.00	–
AR(1)	0.79	−0.00	0.988
Ridge	0.81	−0.05	0.136
Random forest	0.84	−0.13	0.075
OPFR-Net small fixed	1.06 ± 0.10	-0.83 ± 0.34	0.031
OPFR-Net tuned	0.91 ± 0.05	-0.32 ± 0.14	0.009

Table 5. Next-month risk classification (test prevalence: solvency risk 0.91, liquidity risk 0.28). Neural-network rows are means over five seeds.

Target / model	Accuracy	Bal. acc.	AUC	Brier
Solvency risk (FR¡1)				
Persistence	0.93	0.78	0.78	0.07
Majority class	0.91	0.50	–	0.09
Logistic regression	0.89	0.58	0.82	0.06
Random forest	0.88	0.48	0.84	0.06
Standard DNN	0.84	0.46	0.40	0.13
OPFR-Net (paper config)	0.84	0.50	0.47	0.15
OPFR-Net tuned	0.81	0.44	0.29	0.16
Liquidity risk (cover¡1.5m)				
Persistence	0.74	0.66	0.66	0.26
Majority class	0.72	0.50	–	0.28
Logistic regression	0.72	0.56	0.64	0.21
Random forest	0.70	0.60	0.66	0.20
Standard DNN	0.68	0.54	0.55	0.22
OPFR-Net (paper config)	0.61	0.47	0.49	0.24
OPFR-Net tuned	0.65	0.49	0.52	0.23

all five seeds, but the seed standard deviation of the bond-share effect (0.083 pp) is as large as its mean. The equity share (+0.024 pp) is positive in four seeds; every other variable or group is at most 0.011 pp, changes sign across seeds, or is negative. Within a single seed the sampling standard deviation over permutations (0.01–0.11 pp) is of the same order as the means. Scenario effects (Table 7) are tiny: the largest is $+0.035$ pp for a 10% rise of the government-bond share, i.e. 4% of the forecast error, and the benefit-outflow scenario changes sign between seeds. The contribution scenario has the counter-intuitive sign (a 10% fall in contributions raises the forecast funding ratio, by 0.014 pp, in all seeds); since this is 1.6% of the forecast error and the network does not beat persistence, we regard it as an artefact of a weakly identified model, not as an actuarial finding. Because the tuned network does not beat persistence, these small effects show what the network uses, not that the variables carry predictive information; rankings of risk drivers from these models should not be interpreted as actuarial evidence.

5.5. Positive control (RQ5)

When a known signal is planted in the same features, the same protocol finds it: persistence RMSE rises to 1.02 pp, ridge, random forest and gradient boosting reach 0.50, 0.48 and 0.38 pp, and the tuned OPFR-Net reaches

Table 6. Permutation importance of the tuned OPFR-Net (increase in out-of-sample RMSE, pp; mean over five seeds).

Variable or group	ΔRMSE (pp)	SD over seeds	Seeds > 0
Group: Portfolio mix	+0.077	0.035	5/5
sbn_share	+0.074	0.083	5/5
saham_share	+0.024	0.021	4/5
funding_ratio	+0.011	0.013	4/5
Group: Funding ratio (lag)	+0.010	0.017	4/5
return_bulanan	+0.005	0.031	2/5
Group: Return & volatility	+0.002	0.032	2/5
manfaat_rate	−0.001	0.011	2/5
deposito_share	−0.002	0.014	2/5
volatilitas_6bln	−0.003	0.006	1/5
Group: Liquidity cover	−0.018	0.029	2/5
likuiditas_cover	−0.018	0.028	2/5
iuran_rate	−0.039	0.036	1/5
Group: Contribution & benefit flows	−0.043	0.022	0/5

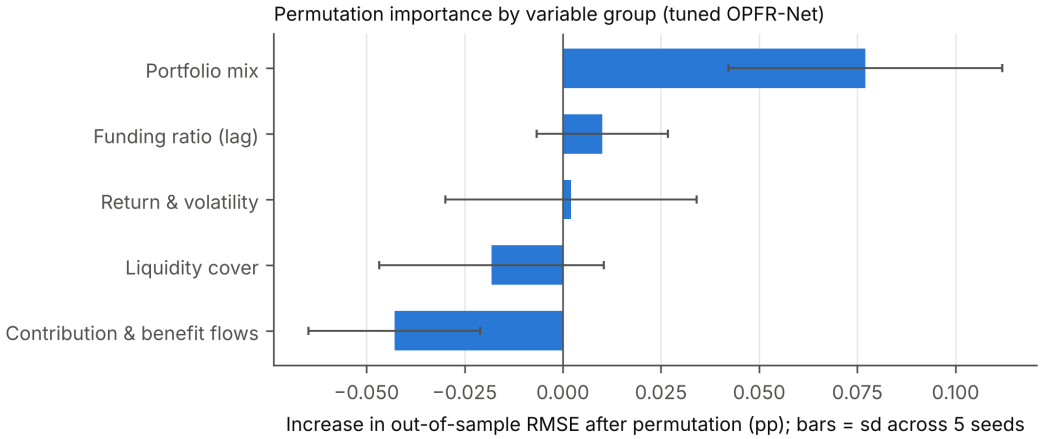


Figure 4. Permutation importance with seed variability (tuned OPFR-Net).

Table 7. Scenario sensitivity of the forecast funding ratio (tuned OPFR-Net).

Scenario	Mean effect on FR forecast (pp)	SD over seeds	Same sign, all seeds	|effect|/RMSE
Investment return −10%	−0.019	0.006	yes	2.2
Volatility +10%	−0.015	0.002	yes	1.7
Contribution inflow −10%	+0.014	0.005	yes	1.6
Benefit outflow +10%	−0.004	0.005	no	0.4
Liquid cover −10%	+0.015	0.012	yes	1.8
Government-bond share +10%	+0.035	0.017	yes	4.0
Lagged funding ratio −1 pp	+0.034	0.009	yes	3.8

0.64 ± 0.10 pp (MSE skill 0.60 ± 0.13 relative to persistence, DM $p < 0.001$). The negative result on the real data is therefore not a coding artefact. The control also illustrates that tree ensembles, not the networks, are the most accurate once a signal exists.

6. Discussion

For monthly industry-level monitoring of Indonesian DB funds, the most informative forecast of next month's funding ratio is this month's funding ratio. A supervisor would not gain accuracy from the networks tested here, and there is a risk of false reassurance if an explainability ranking is read as a causal list of risk drivers. The framework is still useful as an audit tool: any proposed model must beat persistence and AR(1) under a leakage-free protocol, with seed variability and a positive control reported.

Plausible reasons, which we have not tested separately, are (i) the series are short (about 100 rows) relative to the number of parameters, (ii) the funding ratio is strongly persistent, so one-month changes are small and noisy, (iii) the available predictors describe flows and portfolio composition but not the actuarial revaluation of liabilities or market valuation changes at fund level, and (iv) industry aggregation removes cross-sectional variation that a network could exploit. The tuned network was worse than a small fixed network, which is consistent with the weak validation signal.

Limitations. (a) Industry-level aggregates, one forecast horizon (one month) and 57 test months limit power; (b) the thresholds for solvency and liquidity risk are assumptions, and solvency risk is highly imbalanced in the test period; (c) macro variables cover only 2020 onwards; (d) the original manuscript's liability series could not be traced to a public source and was not used; (e) we use permutation importance, not SHAP values, and do not claim causal interpretation. **Future work.** Panel data at fund level, longer pre-specified horizons (3, 6 and 12 months, with multiple-testing control), sequence models, and richer actuarial inputs are the natural directions in which deep learning could be given a fairer chance; any such analysis should be specified before looking at test results.

7. Conclusion

We developed OPFR-Net, a multi-output neural framework with a nested hyper-parameter search, and evaluated it on a documented monthly OJK dataset of Indonesian DB pension funds under a leakage-free expanding-window protocol. No learned model beat persistence for next-month funding ratio; the standard DNN and the original OPFR-Net configuration were significantly worse. Macroeconomic variables did not help, the risk heads did not improve on label persistence, and permutation importance and scenario effects were small and unstable across seeds. A positive control confirmed that the pipeline detects a real signal. The practical message is a methodological one: pension-risk deep learning claims should be supported by benchmarks, chronological evaluation, tests and open data, and the data and code released with this paper allow other researchers to extend the analysis with richer data.

Data and code availability

The computational code used to develop and evaluate OPFR-Net is publicly available in the GitHub repository at https://github.com/jiunsasmita/fundriskdnn. The repository provides the Python implementation of the proposed multi-output neural network and the computational code supporting the analyses presented in this study. These implementations are intended to facilitate the replication and verification of the model evaluation and forecasting results reported in this paper. The empirical data used in the analysis were obtained from publicly available statistical sources, including pension fund statistics from the Financial Services Authority of Indonesia (OJK), the BI 7-Day Reverse Repo Rate from Bank Indonesia, and consumer price inflation data from Statistics Indonesia (BPS). The availability of the computational code supports research transparency and provides a basis for further methodological developments and independent assessment of the proposed framework.

Acknowledgement

This research was funded by the Research and Innovation for Advanced Indonesia (RIIM) programme, facilitated by the National Research and Innovation Agency (BRIN), through its Deputy for Research and Innovation Facilitation, in collaboration with the Education Fund Management Agency (LPDP), under grant/contract numbers 52/IV/KS/03/2026 and 001/DST/IT13.D/U.03/2026. The authors gratefully acknowledge BRIN and LPDP for their financial support and commitment to advancing research, innovation, science, and technology in Indonesia.

REFERENCES

1. C. I. Nkeki, *Stochastic funding of a defined contribution pension plan with proportional administrative costs and taxation under mean-variance optimization approach*, Statistics, Optimization and Information Computing, vol. 2, no. 4, pp. 323–338, 2014.
2. A. Yaakoub, H. Oukhouya, M. Elhia, T. Zari, and R. Guerbaz, *Deep learning for financial time series: Does LSTM outperform ARIMA and SVR in international stock market predictions?*, Statistics, Optimization and Information Computing, vol. 14, pp. 2181–2199, 2025. doi:10.19139/soic-2310-5070-2602.
3. C. Bekkaye, H. Oukhouya, T. Zari, R. Guerbaz, and H. El Bouanani, *Advanced strategies for predicting and managing auto insurance claims using machine learning models*, Statistics, Optimization and Information Computing, vol. 14, no. 3, 2025.
4. W. Ekasasmita, K. Tunnisa, and M. T. Aditya, *Forecasting nonstationary time series based on discrete Hilbert transform*, Statistics, Optimization and Information Computing, vol. 13, no. 1, pp. 368–377, 2024. doi:10.19139/soic-2310-5070-2060.
5. R. Richman, *AI in actuarial science – a review of recent advances – part 1*, Annals of Actuarial Science, vol. 15, no. 2, pp. 207–229, 2021.
6. R. Richman, *AI in actuarial science – a review of recent advances – part 2*, Annals of Actuarial Science, vol. 15, no. 2, pp. 230–258, 2021.
7. M. V. Wüthrich, and M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer, 2023. doi:10.1007/978-3-031-12409-9.
8. G. Iyengar, and A. K. C. Ma, *A robust optimization approach to pension fund management*, Journal of Asset Management, vol. 11, pp. 163–177, 2010.
9. D. Hainaut, and G. Deelstra, *Optimal funding of defined benefit pension plans*, Journal of Pension Economics and Finance, vol. 10, no. 1, 2011. doi:10.1017/S1474747210000016.
10. F. Menoncin, and E. Vigna, *Mean-variance target-based optimisation for defined contribution pension schemes in a stochastic framework*, Insurance: Mathematics and Economics, 2017.
11. Q. Zhao, Y. Shen, and J. Wei, *Mean-variance investment and contribution decisions for defined benefit pension plans in a stochastic framework*, Journal of Industrial and Management Optimization, vol. 17, no. 3, 2021.
12. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, *Statistical and Machine Learning forecasting methods: Concerns and ways forward*, PLoS ONE, vol. 13, no. 3, e0194889, 2018. doi:10.1371/journal.pone.0194889.
13. H. Hewamalage, K. Ackermann, and C. Bergmeir, *Forecast evaluation for data scientists: common pitfalls and best practices*, Data Mining and Knowledge Discovery, vol. 37, pp. 788–832, 2023.
14. L. Grinsztajn, E. Oyallon, and G. Varoquaux, *Why do tree-based models still outperform deep learning on typical tabular data?*, Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track.
15. L. J. Tashman, *Out-of-sample tests of forecasting accuracy: an analysis and review*, International Journal of Forecasting, vol. 16, no. 4, pp. 437–450, 2000.
16. F. X. Diebold, and R. S. Mariano, *Comparing predictive accuracy*, Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 253–263, 1995.
17. D. Harvey, S. Leybourne, and P. Newbold, *Testing the equality of prediction mean squared errors*, International Journal of Forecasting, vol. 13, no. 2, pp. 281–291, 1997.
18. J. Bergstra, and Y. Bengio, *Random search for hyper-parameter optimization*, Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
19. D. P. Kingma, and J. Ba, *Adam: A method for stochastic optimization*, International Conference on Learning Representations (ICLR), 2015.
20. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: a simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.
21. D. Hendrycks, and K. Gimpel, *Gaussian error linear units (GELUs)*, arXiv:1606.08415, 2016.
22. L. Breiman, *Random forests*, Machine Learning, vol. 45, pp. 5–32, 2001.
23. J. H. Friedman, *Greedy function approximation: a gradient boosting machine*, Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
24. A. E. Hoerl, and R. W. Kennard, *Ridge regression: biased estimation for nonorthogonal problems*, Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
25. A. Fisher, C. Rudin, and F. Dominici, *All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously*, Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019.
26. S. M. Lundberg, and S.-I. Lee, *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems 30, pp. 4765–4774, 2017.
27. Otoritas Jasa Keuangan (OJK), *Statistik Dana Pensiun (pension fund statistics: net assets, actuarial liabilities, investment portfolio, contributions, benefit payments)*, monthly series, data snapshot 28 May 2026, https://www.ojk.go.id.

28. Bank Indonesia, *BI 7-Day (Reverse) Repo Rate*, https://www.bi.go.id.
29. Statistics Indonesia (BPS), *Consumer price inflation*, https://www.bps.go.id.