JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e–ISSN: 2614–8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 22, No. 2, January 2026, pp. 511-526
DOI: 10.20956/j.v22i2.48609

Random Forest vs Elastic-Net Penalized Logistic Regression for Patient Discharge Classification in BPJS Primary Care

David Kevin Hutabarat[1*], Sawaluddin[2] , Syahriol Sitorus[3], Sutarman[4]
[1,2,3,4]Department of Mathematics, Universitas Sumatera Utara
***Email**: [1]davidkevin1404@gmail.com, [2]sawal@usu.ac.id, [3]syahriol@usu.ac.id, [4]sutarman@usu.ac.id*
**Corresponding Author*

Abstract

This study analyzes and compares Random Forest and Penalized Logistic Regression (Elastic Net, SAGA solver) for classifying patient discharge status at BPJS Kesehatan primary care facilities (FKTP). The large-scale dataset consists entirely of nominal predictors with class imbalance (~64.9% majority). The experimental design applies an 80/20 train–test split, one-hot encoding, and class_weight = balanced for both models. Hyperparameters are tuned via a staged coarse→fine randomized search with a local-optimum convergence rule (improvement threshold $\varepsilon = 1e{-}6$, patience = 10), followed by 10-fold cross-validation for internal validation and final testing on the hold-out set. We evaluate three primary metrics: F1-Score, Precision–Recall AUC (PR-AUC), and Brier Score. On the test set, Random Forest attains F1 = 0.996679, PR-AUC = 0.999933, and Brier = 0.002646; Penalized Logistic Regression attains F1 = 0.996676, PR-AUC = 0.999928, and Brier = 0.002017. The near-identical F1 and PR-AUC indicate comparable discrimination between methods, while the lower Brier Score for Penalized Logistic Regression demonstrates superior probability calibration. Overall, both approaches lie on the same performance plateau for discrimination, with a consistent calibration advantage for Penalized Logistic Regression; method choice can thus be guided by whether operational needs prioritize calibrated probabilities or flexible non-linear decision boundaries.

Keywords: BPJS Kesehatan, Elastic Net, Hyperparameter Tuning, Imbalanced Classification, Random Forest

1. INTRODUCTION AND PRELIMINARIES

Primary-care information systems now record millions of patient encounters each year, mostly composed of categorical data such as administrative codes, service types, and facility identifiers. These categorical-only datasets challenge conventional predictive modeling due to high-cardinality encoding, correlation between variables, and class imbalance typical of healthcare outcomes. In Indonesia's BPJS Kesehatan primary care (FKTP), discharge status serves as a critical operational

target variable for optimizing referral control and patient flow management [24]. Therefore, predictive models in this context must be scalable, interpretable, and capable of handling categorical variables while maintaining calibration quality.

Random Forest (RF) and Elastic-Net–penalized logistic regression (PLR-EN) represent two contrasting yet complementary paradigms. RF, an ensemble of decision trees built on bootstrap samples and random feature subsets, excels in capturing nonlinearities and interactions [4],[36]. However, multiple recent analyses emphasize that impurity-based importance can distort interpretability when applied to categorical predictors with unequal cardinalities, a recurrent issue in electronic health record (EHR) data [7],[14]. In response, conditional or permutation importance, and calibration extensions such as Califorest, have been developed to improve probability reliability [34].

PLR-EN, on the other hand, retains the interpretability and probabilistic consistency of logistic regression while addressing multicollinearity and overfitting through combined ℓ_1–ℓ_2 penalties. Studies in chronic disease modeling and cost prediction show that elastic-net regularization improves both discrimination and calibration when applied to high-dimensional one-hot–encoded administrative datasets [1],[10],[16]. Furthermore, penalized logistic frameworks are often favored in operational healthcare analytics for their transparent coefficients and ease of model auditing—key for institutional deployment [39].

However, empirical comparisons between these paradigms yield inconsistent findings. Large-scale benchmarks confirm RF's superior discrimination in many general-purpose tabular datasets [44], yet systematic clinical reviews have shown that machine learning does not consistently outperform penalized logistic regression once study quality, calibration, and validation rigor are controlled [4],[5]. Recent meta-analyses highlight that calibration drift remains a significant limitation for unadjusted RF models in longitudinal healthcare applications [10].

Context-specific results are emerging. In hospital utilization and readmission prediction, RF often outperforms linear models in AUC but shows poorer probability calibration unless isotonic or Platt scaling is applied [12],[15]. Conversely, Elastic-Net retains calibration superiority, especially under class imbalance and categorical heterogeneity [6],[13]. This tradeoff between nonlinear discrimination and probabilistic calibration is now recognized as central in healthcare predictive modeling [42].

Despite the growing literature, two specific gaps persist in the context of national-scale, categorical-only healthcare systems like BPJS: (i) lack of head-to-head evidence comparing RF and PLR-EN on ultra-large (millions of records) administrative datasets with exclusively nominal predictors. Most published comparisons involve mixed or continuous variables, or limited sample sizes (<100k) [1],[11],[13]; and (ii) evaluation metrics remain ROC-centric, even though imbalanced operational outcomes (e.g., rare adverse discharges) demand precision–recall and probabilistic calibration analyses for real-world decision support [6],[16].

Accordingly, this study addresses those gaps by conducting a head-to-head evaluation of Random Forest (RF) and Elastic-Net–penalized logistic regression (PLR-EN) on >4 million BPJS-FKTP encounters with exclusively nominal predictors in an imbalanced setting. We employ a unified, compute-aware pipeline—full one-hot encoding, explicit class-imbalance handling, consistent data partitioning/validation, and hyperparameter tuning specified for reproducibility—and we jointly assess discrimination and probabilistic accuracy using F1-Score, Precision–Recall AUC, and Brier Score. The contribution is twofold: (i) large-scale, all-nominal evidence from a national primary-care system that clarifies when ensemble nonlinearity outperforms (or is matched by) regularized linear decision boundaries; and (ii) transparent modeling guidance on design choices (encoding, categorical cardinality, and regularization) that materially influence operational deployment in primary-care analytics.

2. MATERIAL AND METHODS

2.1 Random Forest

Decision-tree methods such as CART partition the feature space recursively using impurity criteria (e.g., the Gini index) until a stopping rule is met; the Gini at node$t$ can be written $i(t) = 1 - \sum_{k=1}^{K} p(k|t)^2$and split quality is evaluated by the (weighted) decrease in impurity across children. While CART is flexible and interpretable, a single tree is deterministic and high-variance—small data perturbations can change the structure markedly—so it is prone to overfitting unless complexity is constrained (e.g., via cost-complexity pruning) [5],[50].

Random Forest (RF) reduces variance and stabilizes predictions by ensembling many de-correlated CARTs trained on bootstrap samples (bagging) and using random feature subsets at each split (random subspace). The forest predicts by majority vote

$$\arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{2.1}$$

and yields class probabilities by averaging leaf-level proportions

$$\hat{p}(x) = \frac{1}{T}\sum_{t=1}^{T} p_t(x), \qquad p_t(x) = \frac{count_{positive}^{(t)}\big(leaf(x)\big)}{count_{positive}^{(t)}\big(leaf(x)\big) + count_{negative}^{(t)}\big(leaf(x)\big)} \tag{2.2}$$

These two randomness sources de-correlate trees and lower ensemble variance, while probability averaging improves stability across trees [4],[25].

For categorical, one-hot encoded (OHE) tabular data, RF's impurity-based splitting naturally handles multi-level predictors and captures non-linear interactions. In practice, the number of candidate features per split is set to a small random subset—often$\sqrt{p}$ or $\log_2 p$—to enhance tree diversity.

RF supports variable importance summaries via impurity accounting [21]. The mean decrease in impurity (MDI) for feature$j$sums the weighted impurity drops across all splits using$j$, typically normalized across features:

$$\Delta I_s = w(t)G(t) - w(t_L)G(t_L) - w(t_R)G(t_R),$$
$$Imp_{MDI}(j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S_t:feat(s)=j} \Delta I_s \tag{2.3}$$

Because OHE creates many indicators per source variable, raw MDI at the indicator-level can be mechanically biased toward high-cardinality predictors; therefore, we aggregate OHE-level importances back to the source variable, and uses OHE-level lists only for drill-down [33],[45].

Key hyperparameters govern bias–variance and inter-tree correlation: (i) T(number of trees) improves stability to a plateau; (ii) maximum depth$d$and minimum split/leaf sizes(m_s, m_ℓ)control base-tree variance; (iii) the per-split feature count $\phi \in \{\sqrt{p}, \log_2 p\}$ de-correlates trees [36].These choices are reflected explicitly in the search grids and in the observed plateaus under the coarse→fine procedure.

Finally, for imbalanced classification, RF can incorporate class weights during training [9]. In the pipeline, class weighting affects both impurity computation and terminal-leaf probability estimates within cross-validation and model refits—ensuring that split selection and probability aggregation respect the skewed class distribution. This makes RF a principled, scalable option for national-scale, all-categorical administrative data, and a meaningful comparator to penalized logistic regression in terms of both discrimination and probability quality.

2.2 Penalized Logistic Regression (Elastic Net)

Binary logistic regression models the conditional probability $\pi(x) = \Pr(Y = 1|x)$through the logit link, with categorical predictors commonly represented by one-hot encoding (OHE) so that each coefficient corresponds to a level's log-odds difference from a chosen baseline. Under class imbalance, a weighted negative log-likelihood is often minimized so that parameter estimates reflect the intended operating distribution [27].

To improve generalization, penalized logistic regression augments the loss with a penalty on coefficients. The Elastic Net penalty combines ℓ_1 (LASSO) and ℓ_2 (Ridge) regularization in a single convex objective,

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n} w_i[-y_i \log \pi_i - (1-y_i)\log(1-\pi_i)] + \lambda\left(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\right), \qquad (2.4)$$

$$\pi_i = \sigma\left(\beta_0 + x_i^{\top}\beta\right)$$

Where $\lambda > 0$ controls total shrinkage and $\alpha \in [0,1]$ mixes ℓ_1 and $\ell_2$components [55]. The construction inherits sparsity from LASSO and stability/grouping from Ridge, making it well-suited to high-dimensional, correlated OHE designs

Historically, Ridge (ℓ_2) reduces variance but does not select variables [26], whereas LASSO (ℓ_1) performs automatic selection but can be unstable under strong predictor correlation [49]. Elastic Net alleviates both limitations by combining their effects. In applications with many categorical levels, Elastic Net's joint shrinkage–selection retains groups of correlated indicators ("grouping effect") and yields more stable estimates than LASSO alone [47].

Efficient optimization is critical at scale. The SAGA algorithm—a variance-reduced stochastic gradient method for composite (smooth + nonsmooth) objectives—provides unbiased gradient estimates, supports proximal updates for ℓ_1, and enjoys linear convergence in strongly convex settings while remaining memory-efficient on sparse, high-dimensional data [14],[40]. In proximal form, SAGA computes a gradient step on the smooth (log-loss) part and applies soft-thresholding to handle the $\ell_1$term, enabling fast, stable training of Elastic-Net logistic regression on large OHE matrices.

Key hyperparameters govern Elastic Net's behavior: the inverse regularization $C = 1/\lambda$ (capacity), the mixing parameter $\text{l1\_ratio} = \alpha$ (sparsity vs. grouping), and the solver's iteration budget $\text{max\_iter} = it$. A large C weakens regularization (risking overfit), whereas smaller C strengthens shrinkage; $\alpha = 1$ reduces to LASSO and $\alpha = 0$ to Ridge. With class weighting, the fitted intercept corresponds to a weighted baseline (pseudo-intercept), and the optimized objective becomes the weighted log-loss plus penalty. For interpretability, logistic coefficients exponentiate to odds ratios, providing effect sizes directly on the multiplicative odds scale [27].

2.3 Hyperparameter Optimization (Coarse→Fine with Local Optimum)

Hyperparameter optimization (HPO) can be framed as black-box minimization of a noisy generalization-error estimator $f(\theta)$ over a mixed search space Θ (continuous, integer, and categorical coordinates). Each function evaluation entails training a model with configuration θ and scoring it via a resampling estimator, typically stratified K-fold cross-validation [18]. Because $f(\theta)$ is expensive and stochastic, effective HPO balances exploration (broad coverage of Θ) and exploitation (refinement near promising regions) while controlling variance and avoiding selection bias [15].

The two-stage coarse→fine procedure implemented here—broad randomized exploration followed by localized refinement—mirrors recent advances in cost-efficient hyperparameter optimization, where multi-fidelity and bandit-based frameworks progressively allocate resources from coarse-grained to fine-grained evaluations. These approaches reduce redundant trials in low-

utility regions and focus computational effort on promising configurations, achieving substantial efficiency gains in large-scale model tuning [52],[54]. The two stage procedure:

(i) Coarse stage (global exploration). Use randomized or sparse grid sampling over wide, preferably log-scaled ranges for scale parameters (e.g. C in penalized logistic regression; number of trees in Random Forest), and broad categorical choices (e.g., max_features). The goal is to map the landscape, detect robust basins of good performance, and identify multiple incumbents rather than a single brittle winner [2].

(ii) Fine stage (local refinement). Choose the strongest incumbent $\theta^{\star}$ from the coarse stage and define a neighborhood $\mathcal{N}_{\rho}(\theta^{\star})$ with per-parameter radii ρ adapted to the parameter's scale: log-additive steps for positive reals (e.g., $C \to C \times \{0.5,0.8,1,1.25,2\}$), small absolute steps for bounded reals (e.g., ± 0.05 within $[0,1]$), ± 1 or small sets for integers (e.g., max_depth), and short, curated sets for categoricals (e.g., max_features in $\{\sqrt{p}, \log_2 p\}$) [15].

Sample candidates within $\mathcal{N}_{\rho}(\theta^{\star})$ optionally shrinking ρ geometrically after each improvement (trust-region behavior).

Table 2.1. Coarse and Fine Stage Search Space for Random Forest

Hyperparameter (θ)	Coarse Stage Search Space	Fine Stage Search Space
n_estimators (T)	300, 500, 700, 1000	400, 500, 600
max_depth (d)	20, 30, 40, 50	20
max_features (ϕ)	sqrt, log2	log2
min_samples_split (m_s)	5, 10	5, 10
min_samples_leaf (m_ℓ)	2, 4, 8	6, 8,10

Note for Random Forest: Criterion: Gini; class_weight="balanced"; fixed random_state=42; $p$denotes the number of features after one-hot encoding.

Table 2.2. Coarse and Fine Stage Search Space for Penalized Logistic Regression (Elastic Net)

Hyperparameter (θ)	Coarse Stage Search Space	Fine Stage Search Space
C	0.001, 0.01, 0.1, 1, 10	0.08, 0.09, 0.10, 0.11, 0.12
L1_ratio (α)	0.1–0.9 (step 0.1)	0.66, 0.70, 0.74
max_iter (it)	300, 500, 800, 1000, 1200	500

Note for Penalized Logistic Regression: penalty="elasticnet", solver="saga", class_weight="balanced", fixed random_state=42.

A configuration $\theta^{\star}$ is considered (ε, ρ)-local-optimal if no sampled neighbor $\tilde{\theta} \in \mathcal{N}_{\rho}(\theta^{\star})$ improves the objective by more than ε across patience iterations [30], [53]. Practical rules:

(i) Improvement threshold (ε): set to $1 \cdot 10^{-6}$ on the CV F1 (threshold 0.5) objective (used consistently during search).

(ii) Patience (ρ): 10 consecutive non-improving evaluations before stopping the local search.

(iii) Neighborhood verification: when a plateau is detected, draw a final batch of neighbors within the same radius to guard against false plateaus due to CV noise.

(iv) Tie-breakers: prefer simpler configurations (shallower trees, larger regularization) when scores are statistically indistinguishable; as a secondary tie-breaker, prefer the model with lower fold-to-fold variance [16].

Variance control in the objective:

(i) Use stratified 10-fold CV with fixed fold assignments reused across configurations (paired comparisons).

(ii) Aggregate fold scores as mean ±SD (with a median check for robustness); log every trial (configuration, score, seed, wall-time) [28].
(iii) Keep preprocessing inside a single pipeline so encoders fit only on training folds (no leakage).
(iv) Optionally run a confirmatory resampling on top candidates before final selection (budget-permitting).

Post-selection & evaluation:

(i) After identifying a locally optimal configuration, refit on the full 80% training split.
(ii) Optionally calibrate probabilities on an internal validation split if required [8].
(iii) Compute F1 (threshold 0.5), PR-AUC, and Brier Score once on the untouched 20% test split (report uncertainty where feasible)

2.4 Evaluation

Model assessment in imbalanced binary classification should distinguish discrimination (how well a model separates positives from negatives) from probability accuracy/calibration (how close predicted probabilities are to true event risks) [8],[27]. Let $y_i \in \{0,1\}$ and $p_i \in \{0,1\}$ be the predicted probability for case i. For a threshold $\tau = 0.5$, predicted labels are $\hat{y}_i = 1\{p_i \geq \tau\}$ and the confusion-matrix counts are TP, FP, TN, and FN.

(i) F1-Score (Thresholded Discrimination)

Precision and recall are defined as [35]:

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN} \tag{2.5}$$

and the harmonic mean

$$\mathrm{F1} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{2.6}$$

The F1-score summarizes positive-class retrieval at the chosen operating threshold. It is insensitive to true negatives and thus suitable under class imbalance, though it remains threshold-dependent [38].

(ii) Precision–Recall AUC (Threshold-Free Discrimination under Imbalance).

The PR curve plots precision as a function of recall while τ varies from 1 to 0. Its area,

$$\mathrm{PR-AUC} = \int_0^1 \mathrm{Precision}(\mathrm{Recall})\, d(\mathrm{Recall}) \tag{2.7}$$

provides a threshold-free summary that is more informative than ROC-based summaries when positives are rare because it focuses on performance over the positive class [12],[18]. Implementations typically compute Average Precision (AP) as a step-wise integral: $\mathrm{AP} = \sum_k \Delta\mathrm{Recall}_k \cdot \mathrm{Precision}_k^*$ where $\mathrm{Precision}_k^*$ is a nonincreasing interpolation of precision at recall step$k$. It is good practice to report the baseline (class prevalence) line on PR plots for context [3].

(iii) Brier Score (Probability Accuracy and Calibration).

The Brier Score (BS) is the mean squared error between predicted probabilities p_i and actual outcomes [6]:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{2.8}$$

It is a proper scoring rule where lower is better and decomposable as $\mathrm{BS} = \mathrm{REL(reliability)} - \mathrm{RES(resolution)} + \mathrm{UNC(uncertainty)}$ [31]. The baseline score for a constant forecaster predicting class prevalence π is $\mathrm{BS} = \pi(1-\pi)$. Well-calibrated models should improve upon this baseline.

Because BS evaluates the full predictive distribution, it complements F1 and PR-AUC by rewarding well-calibrated probabilities even when thresholded retrieval is similar [7], [22].

To avoid optimistic bias, model selection is performed on resampling splits, and a held-out test set is used once for final reporting [51]. For imbalanced data, stratified splitting preserves prevalence across folds/sets.Finally, jointly reporting F1 (thresholded retrieval), PR-AUC/AP (threshold-free discrimination), and Brier Score (probabilistic calibration) provides a balanced view of model performance under class imbalance and probabilistic decision-making.

2.5 FKTP BPJS Kesehatan (First-Level Health Facilities)

Indonesia's National Health Insurance (Jaminan Kesehatan Nasional, JKN) is administered by BPJS Kesehatan, which ensures population-wide access to essential health services through a hierarchical provider network. First-Level Health Facilities (Fasilitas Kesehatan Tingkat Pertama, FKTP) serve as the primary gatekeepers in this system, providing promotive, preventive, and basic curative services while managing referrals to higher-level care when clinically indicated.

FKTP encompass diverse organizational types—public community health centers (puskesmas), private general-practice clinics, solo general practitioners, and dental practices—reflecting Indonesia's mixed public–private primary-care delivery model [23]. As of June 2025, BPJS Kesehatan reported partnerships with 23,497 FKTP nationwide, comprising 10,134 puskesmas (43.1%), 6,530 primary clinics (27.8%), 4,463 private general practitioners (19.0%), 1,177 dental practices (5.0%), 571 military clinics (2.43%), 570 police clinics (2.43%), and 52 Type D Pratama hospitals (0.22%) [32].

Operationally, FKTP exhibit heterogeneity in ownership, staffing, and service scope. Standard administrative classifications distinguish:

(i) Facility type—such as puskesmas, private clinic, solo GP, dental practice, laboratory, and affiliated network provider;
(ii) Service level—outpatient first-level care (Rawat Jalan Tingkat Pertama, RJTP), inpatient first-level care (Rawat Inap Tingkat Pertama, RITP), and promotive–preventive services; and
(iii) Clinical unit type (poli)—including general, dental, maternal and child health (KIA), emergency, family planning, delivery, laboratory, immunization, nutrition, chronic-disease program, home visit, counseling, and geriatric services.

Within the tiered referral system, FKTP act as the entry point to higher-level care (secondary and tertiary facilities). Referral records typically capture both the referral decision (yes/no) and destination attributes (facility type, ownership, specialization), enabling analytic studies of referral flows and care coordination within JKN's primary-care network [23].

Routine FKTP encounter data collected by BPJS Kesehatan include standardized fields such as facility characteristics, service level, clinical unit, participant segment, visit type (e.g., sick vs. preventive), geographic location, and referral destination. These structured variables provide a comprehensive foundation for statistical modeling and machine-learning analysis of primary-care utilization, continuity, and outcomes [17].

2.6 Data

This study uses secondary encounter records from Indonesia's BPJS Kesehatan First-Level Health Facilities (FKTP) for 2015–2020, obtained from the BPJS Kesehatan Dataset Portal (access here: https://s.id/DatasetFKTP). The dataset comprises 4,056,897 visit-level observations and 11 variables. The binary outcome is Discharge Status (Kelas Status Pulang Peserta), coded 1 = Sehat (recovered and well-visit) and 0 = Belum Sehat (continued outpatient care, external/internal referral, discharge against medical advice, or death). Class balance is imbalanced: 2,634,511 (64.9%) Belum Sehat and 1,422,386 (35.1%) Sehat.

Predictors consist of 10 FKTP administrative fields, all nominal: Jenis FKTP, Tipe FKTP, Tingkat Pelayanan FKTP, Jenis Poli FKTP, Segmen Peserta FKTP, Jenis Kunjungan FKTP, Kepemilikan FKTP, Jenis Faskes Tujuan Rujukan, TipeFaskes Tujuan Rujukan, and Provinsi FKTP (Table 2.3).

For modeling, data were split into 80% training and 20% testing with stratification by outcome; the effective training size used in derivations is 3,245,517 (implying 811,380 test cases). Categorical predictors were one-hot encoded prior to learning.

Table 2.3. Variables, Descriptions, and Cardinalities (Nominal Predictors)

Variable	Categories	Description
X_1 Jenis FKTP	7	High-level facility category at the first level of care (e.g., puskesmas, primary clinic, solo GP, dental clinic, network/affiliated unit, lab, other first-line units).
X_2 Tipe FKTP	12	Structural subtype used by BPJS for FKTP classification (e.g., stand-alone vs. network/affiliated facility; administrative subtype within FKTP classes).
X_3 Tingkat Pelayanan FKTP	3	Service level provided at FKTP (e.g., RJTP outpatient first-level care; RITP inpatient first-level care; promotive/preventive services).
X_4 Jenis Poli FKTP	28	Clinical unit / polyclinic type available at the facility (e.g., general clinic, dental & oral, maternal–child health, emergency unit, family planning, delivery, lab, immunization, nutrition, chronic programs, home visit, counseling, geriatric, etc.).
X_5 Segmen Peserta FKTP	5	Participant membership segment in JKN (e.g., wage earner, self-employed, government-subsidized/beneficiary, non-wage worker, other statutory segments).
X_6 Jenis Kunjungan FKTP	2	Visit type at FKTP (e.g., sick visit vs. well/maintenance visit; administrative categorization of encounter intent).
X_7 Kepemilikan FKTP	9	Ownership of the facility (e.g., central/local government, state-owned enterprise, private/independent, foundation, military/police, university/other).
X_8 Jenis Faskes Tujuan Rujukan	5	High-level type of referral destination when the encounter is referred onward (e.g., hospital, clinic, puskesmas, other; includes "not referred" coding).
X_9 TipeFaskes Tujuan Rujukan	40	Subtype of referral facility (e.g., general hospital, specialty hospital categories,

Variable	Categories	Description
		maternity/child hospital, specialized clinics, and other coded subtypes).
X_{10} Provinsi FKTP	34	Province of the FKTP location (administrative province code; national coverage).

Table 2.4. Outcome Definition and Class Distribution

Outcome(Y) Label	Definition	Count	Proportion
0 = Belum Sehat	Outpatient follow-up, external/internal referral, discharge against advice, or death	2,634,511	64.9%
1 = Sehat	Recovered and well-visit	1,422,386	35.1%

Notes. Variable definitions (category lists) follow the administrative FKTP taxonomy used in the source data (e.g., 28 polyclinic types; 40 referral facility subtypes). These nominal structures motivate the exclusive use of categorical encodings in subsequent models.

3. RESULTS

3.1 Model Summary

a. Coarse Stage

Under stratified 10-fold CV (fixed folds) with F1 ($\tau = 0.5$) as the objective, the coarse search mapped each model's high-performing basin before any local refinement. Two consistent patterns emerged.

(i) Random Forest (RF): F1 gains plateaued beyond mid-range tree counts; shallower trees (depth ≈ 20) with non-trivial leaf sizes (≈ 6–10) remained competitive when paired with $\log_2$feature subsetting, indicating effective variance control without harming positive-class retrieval.

(ii) Elastic-Net Logistic Regression (SAGA): a moderate capacity regime ($C \approx 0.08$–0.12) with mixed sparsity–stability ($\alpha \approx 0.66$–0.74) yielded a stable F1 plateau, and $it = 500$ sufficed for convergence at those settings. When coarse-stage candidates were statistically indistinguishable, ties were resolved in favor of simpler configurations (e.g., shallower RF trees; stronger PLR-EN regularization). These coarse patterns defined the incumbent regions that seeded the fine (local) search.

Table 3.1. Coarse-Stage Findings

Model	Coarse-Stage Pattern Observed	Incumbent Region Taken to Fine Stage
Random Forest	F1 saturates beyond mid-range T; de-correlated splits ($\log_2 p$) + regularized tree shape (shallower depth, non-trivial leaves)	$T \in \{400,500,600\}$; $d = 20$;$m_s \in \{5,10\}$; $m_\ell \in \{6,8,10\}$; $\phi = \log_2 p$
Elastic-Net (SAGA)	Stable F1 plateau at moderate capacity with mixed ℓ_1–ℓ_2; 500 iterations sufficient	$C \in \{0.08, 0.09, 0.10, 0.11, 0.12\}$;$\alpha \in \{0.66, 0.70, 0.74\}$; $it = 500$

Note: $p$denotes the number of features after one-hot encoding.

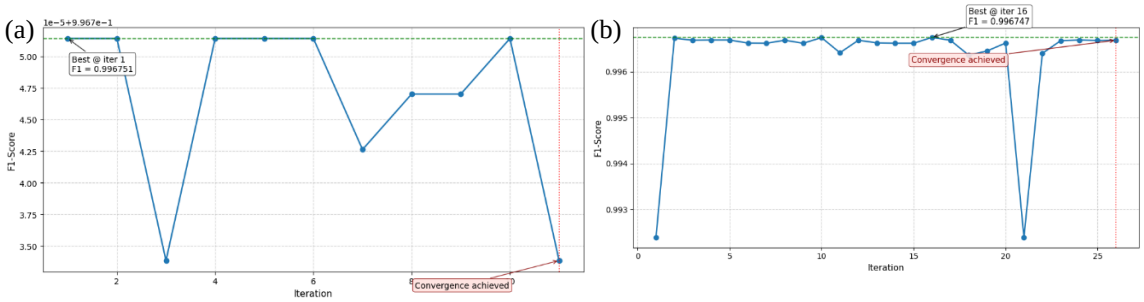

Figure 3.1. (a) Coarse-Stage RF F1 trajectories; (b) Coarse-Stage PLR-EN F1 trajectories.

The coarse stage (Figure 3.1) achieved its goal of identifying robust basins rather than over-committing to a single configuration. For RF, the evidence favors $\log_2$feature subsetting to de-correlate trees and controlled tree complexity to suppress variance, with no material F1 benefit from very deep trees or very small leaves. For PLR-EN, the $C \approx 0.1$ and$\alpha \approx 0.7$ regime balances sparsity (selection of informative one-hot indicators) and stability (grouping under correlated dummies), producing flat, reproducible F1 across folds—an ideal anchor for the subsequent fine search.

b. Fine Stage

Building on the coarse-stage incumbents, the fine stage restricted the search to a local neighborhood and refined configurations with scale-aware steps, under stratified 10-fold CV (fixed folds) and F1 ($\tau = 0.5$)as the sole objective. Two clear optima emerged:

(i) Random Forest (RF). Within the neighborhood $T \in \{400,500,600\}$; $d \approx 20$; $m_s \in \{5,10\}$;$m_\ell \in \{6,8,10\}$; $\phi = \log_2 p$, the best CV F1 was attained at $T = 500$; $d = 20$; $m_s = 10$; $m_\ell = 8$; $\phi = \log_2 p$, with extremely small fold-to-fold dispersion. Neighboring settingsproduced no CV-F1 improvement $\geq \varepsilon$, produced no CV-F1 improvement.

(ii) Penalized Logistic Regression (Elastic Net, SAGA). In the neighborhood $\in \{0.08, 0.09, 0.10, 0.11, 0.12\}$; $\alpha \in \{0.66, 0.70, 0.74\}$; $it = 500$, the optimum was $C = 0.08, \alpha = 0.74, it = 500$. Neighboring settings yielded no CV-F1 improvement $\geq \varepsilon$, and the fold-level scores were stable across repeats.

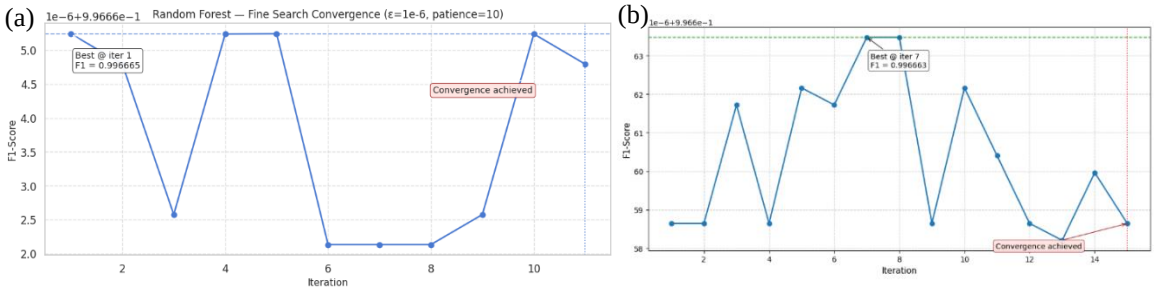

Figure 3.2. (a) Fine-Stage RF F1 trajectories; (b) Fine-Stage PLR-EN F1 trajectories.

Local optimality was verified using the pre-specified criterion: a candidate is (ε, ρ)-local-optimalif no neighbor within the current radius improves the objective by more than $\varepsilon = 10^{-6}$ across 10 consecutive evaluations ($\rho = 10$). For both RF and Elastic-Net:

(i) Consecutive non-improvement: the best incumbent remained unchanged for ≥10 neighbor trials with $\Delta F1 < \varepsilon$ on each trial.

(ii) Neighborhood verification: an additional batch of neighbors sampled at the same radius produced no $\Delta F1 \geq \varepsilon$.
(iii) Parsimony tie-break: where minuscule differences existed, the simpler model (shallower trees/ stronger regularization) was preferred; this rule did not overturn the incumbent.

These checks, together with Figure 3.2, establish that the selected configurations are local optima within their respective fine-stage neighborhoods.

3.2 Cross-Validation

Stratified 10-fold cross-validation (fixed folds) was applied on the 80% training split with all preprocessing inside the pipeline (full one-hot encoding) and class_weight = "balanced" active for both model families. The primary CV objective was F1 at $\tau = 0.5$, while PR-AUC and Brier Score were computed for completeness. Fold construction and notation follow the standard definition $D = F_1 \cup ... \cup F_{10}$ training on $D \backslash F_k$ and scoring on F_k.

Table 3.2. 10-Fold Cross Validation Results

Model	F1	PR-AUC	Brier Score
Random Forest	0.996667	0.999931	0.002689
Penalized Logistic Regression (Elastic Net)	0.996661	0.999925	0.002026

Discrimination is essentially identical across models on CV: $\Delta F_1 \approx 6 \times 10^{-6}$ and $\Delta PR - AUC \approx 6 \times 10^{-6}$. In contrast, Brier Score is consistently lower for Elastic Net, indicating better probability calibration despite matched thresholded retrieval. For the selected Random Forest, the CV-10 F1 = 0.996667 is numerically indistinguishable from the fine-stage best (0.996665), with differences on the order of 10^{-6}. This stability shows the decision function (averaging $\hat{p}_t(x)$ and thresholding at 0.5) remains on the same performance plateau when moving from the fine-stage search CV to the stricter 10-fold CV.

To guard against spurious plateaus from resampling noise, top-2 configurations in each family were re-scored under 3-fold CV across multiple random seeds; the mean ΔF_1(best – runner up) remained below the preset $\varepsilon = 10^{-6}$, supporting local-optimum convergence observed in the fine stage. Overall, 10-fold CV confirms that both models lie on the same discrimination plateau, while Elastic Net provides superior probability accuracy.

3.3 Held-Out Performance

Evaluation on the untouched 20% test split ($N = 811{,}380$; prevalence $\hat{\pi} \approx 0.350857$) confirms that both models lie on the same discrimination plateau, while Penalized Logistic Regression (Elastic Net) yields better probability accuracy.

Table 3.3. Held-Out Test Performance

Model	F1 ($\tau = 0.5$)	PR-AUC	Brier Score
Random Forest	0.996679	0.999933	0.002646
Penalized Logistic Regression (Elastic Net)	0.996676	0.999928	0.002017

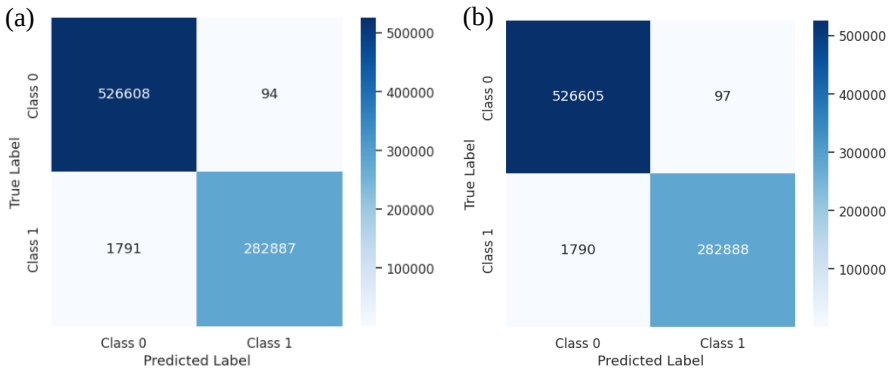

Figure 3.3. (a) RF confusion matrix; (b) PLR-EN confusion matrix

Differences are numerically negligible: $\Delta F_1 \approx 3 \times 10^{-6}$ and $\Delta PR - AUC \approx 5 \times 10^{-6}$, consistent with coarse→fine/CV plateaus. Interpreting the confusion counts (see Figure 3.3) at this scale, Random Forest reduces FP from 97 to 94 (−3) but increases FN from 1,790 to 1,791 (+1), yielding F1 that is effectively identical to Elastic Net.

The Brier Score is materially lower for Elastic Net (0.002017 vs 0.002646), a gap of ≈0.000629 (≈24% lower on the Brier scale). Against the constant-prevalence baseline $\hat{\pi}(1-\hat{\pi}) \approx 0.227756$, both models are excellent, but Elastic Net is consistently better calibrated—coherent with its proper-loss foundation (log-loss) and stable probabilistic outputs—whereas RF probabilities, being leaf-frequency averages, can be slightly under/over-confident in parts of feature space.

On held-out data, discrimination is equivalent for both methods (F1, PR-AUC), while Elastic Net offers a calibration advantage (lower Brier) without sacrificing retrieval, aligning with internal CV findings.

4. CONCLUSION

On a large, all-categorical FKTP dataset, Random Forest and Elastic-Net Penalized Logistic Regression (SAGA) achieve indistinguishable discrimination on the held-out test set (F1 at $\tau = 0.5$ and PR-AUC are effectively equal), while Elastic Net yields superior probability accuracy, lowering the Brier Score by ~0.00063 (≈24% on the Brier scale). A two-stage coarse→fine hyperparameter search with a local-optimum criterion produced stable, reproducible configurations for both models.

Practically, the choice hinges on end-use: when calibrated probabilities are required for risk stratification or cost-sensitive thresholds, Elastic Net (SAGA) is preferred; when interaction capture and ensemble-based variable importance are prioritized, Random Forest remains a robust alternative (with post-hoc calibration if needed). Future work may examine threshold optimization under explicit cost functions, alternative calibration schemes, temporal/external validation to assess generalizability, and comparisons with other categorical-friendly learners.

CONFLICT OF INTEREST

The authors confirm that there is no conflict of interest related to the publication of this article.

REFERENCES

[1] Austin, A. M., Ramkumar, N., Gladders, B., Barnes, J. A., Eid, M. A., Moore, K. O., Feinberg, M. W., Creager, M. A., Bonaca, M., & Goodney, P. P., 2022. Using a cohort study

of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling, *BMC Medical Research Methodology*, Vol. 22, No. 300, 1–10, doi: 10.1186/s12874-022-01774-8.

[2] Bouthillier, X., Laurent, C., & Vincent, P., 2019. Unreproducible Research is Reproducible, in *Proceedings of the 36th International Conference on Machine Learning*, 2019, 725–734. [Online]. Available: https://proceedings.mlr.press/v97/bouthillier19a

[3] Boyd, K., Eng, K. H., & Page, C. D., 2013. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals, in *Joint European conference on machine learning and knowledge discovery in databases*, Berlin: Springer Berlin Heidelberg, 2013, 451–466. doi: 10.1007/978-3-642-40994-3_29.

[4] Breiman, L., 2001. Random Forests, *Machine Learning*, Vol. 45, 5–32, doi: 10.1023/A:1010933404324.

[5] Breiman, L., Jerome Friedman, R.A. Olshen, & Charles J. Stone, 1984. *Classification and Regression Trees*, 1st ed. New York, 1984. doi: 10.1201/9781315139470.

[6] Brier, G. W., 1950. Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, Vol. 78, No. 1, 1–3, doi: 10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

[7] Bröcker, J., 2009. Reliability , Sufficiency , and the Decomposition of, *The Quarterly Journal of the Royal Meteorological Society*, Vol. 135, No. 643, 1512–1519, doi: 10.1002/qj.456.

[8] Calster, B. Van, Mclernon, D. J., Smeden, M. Van, Wynants, L., & Steyerberg, E. W., 2019. Calibration: the Achilles heel of predictive analytics, *BMC Medicine*, Vol. 17, No. 230, doi: 10.1186/s12916-019-1466-7.

[9] Chen, C., Liaw, A., & Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. [Online]. Available: https://statistics.berkeley.edu/tech-reports/666

[10] Chowdhury, M. Z. I., Leung, A. A., Walker, R. L., Sikdar, K. C., Beirne, M. O., Quan, H., & Turin, T. C., 2023. A comparison of machine learning algorithms and traditional regression - based statistical modeling for predicting hypertension incidence in a Canadian population, *Scientific Reports*, Vol. 13, No. 13, doi: 10.1038/s41598-022-27264-x.

[11] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, Y., & Calster, B. Van, 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, *Journal of Clinical Epidemiology*, Vol. 110, 12–22, doi: 10.1016/j.jclinepi.2019.02.004.

[12] Davis, J., & Goadrich, M., 2006. The Relationship Between Precision-Recall and ROC Curves, in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, 233–240. doi: 10.1145/1143844.1143874.

[13] Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., & Matheny, M. E., 2017. Calibration drift in regression and machine learning models for acute kidney injury, *Journal of the American Medical Informatics Association*, Vol. 24, No. 6, 1052–1061, doi: 10.1093/jamia/ocx030.

[14] Defazio, A., Bach, F., & Lacoste-Julien, S., 2014. SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives, *Advances in Neural Information Processing Systems*, Vol. 2, No. January, 1646–1654, doi: 10.48550/arXiv.1407.0202.

[15] Eriksson, D., Pearce, M., Gardner, J. R., Turner, R., & Poloczek, M., 2019. Scalable Global Optimization via Local Bayesian Optimization, in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019, 5496–5507. doi: 10.48550/arXiv.1910.01739.

[16] Falkner, S., Klein, A., & Hutter, F., 2018. BOHB: Robust and Efficient Hyperparameter Optimization at Scale, in *Proceedings of the 35th International Conference on Machine*

Learning, 2018, 1437–1446. [Online]. Available: https://proceedings.mlr.press/v80/falkner18a

[17] Farhan, M., Santosa, B., & Sholihah, M., 2024. Identification of Referral Pattern in Indonesian Primary Healthcare Facilities Using Data Mining Techniques, in *IEEE Technology & Engineering Management Conference - Asia Pacific (TEMSCON-ASPAC)*, 2024, 1–6. doi: 10.1109/TEMSCON-ASPAC62480.2024.11024874.

[18] Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F., 2022. Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning, *Journal of Machine Learning Research*, Vol. 23, 1–61, doi: 10.48550/arXiv.2007.04074.

[19] Flach, P. A., 2015. Precision-Recall-Gain Curves: PR Analysis Done Right, in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015, 838–846. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/2969239.2969333

[20] Galiën, O. P. Van Der, Hoekstra, R. C., Gürgöze, M. T., Manintveld, O. C., Bunt, M. R. Van Den, Veenman, C. J., & Boersma, E., 2021. Prediction of long-term hospitalisation and all-cause mortality in patients with chronic heart failure on Dutch claims data: a machine learning approach, *BMC Medical Informatics and Decision Making*, Vol. 21, No. 303, doi: 10.1186/s12911-021-01657-w.

[21] Genuer, R., Poggi, J.-M., & Tuleau-Malot, C., 2010. Variable selection using random forests, *Pattern Recognition Letters*, Vol. 31, No. 14, 2225–2236, doi: 10.1016/j.patrec.2010.03.014.

[22] Gneiting, T., & Raftery, A. E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, Vol. 102, No. 477, 359–378, doi: 10.1198/016214506000001437.

[23] Handayani, P. W., Dartanto, T., Moeis, F. R., Pinem, A. A., Azzahro, F., Hidayanto, A. N., Denny, & Ayuningtyas, D., 2021. The regional and referral compliance of online healthcare systems by Indonesia National Health Insurance agency and health-seeking behavior in Indonesia, *Heliyon*, Vol. 7, No. 9, doi: 10.1016/j.heliyon.2021.e08068.

[24] Hendrawan, D., Ariawan, I., Sartono, B., Wahyuningsih, W., Negara, S. I., Mawardi, J., Fatah, C. J., Sutara, F. A., & Nugraha, N. S., 2021. Data Sampel BPJS Kesehatan 2015-2020, Jakarta. [Online]. Available: https://data.bpjs-kesehatan.go.id/bpjs-portal/action/blog-detail.cbi?id=79f03774-6397-11ec-bd5e-bb284b79c3ff

[25] Ho, T. K., 1995. Random Decision Forests, in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, 8–12. doi: 10.1109/ICDAR.1995.598994.

[26] Hoerl, A. E., & Kennard, R. W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 12, No. 1, 55–67, doi: 10.2307/1267351.

[27] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X., 2013. *Applied Logistic Regression*, 3rd ed. New Jersey: John Wiley & Sons, Inc., 2013.

[28] Kadra, A., Hutter, F., Lindauer, M., & Grabocka, J., 2021. Well-tuned Simple Nets Excel on Tabular Datasets, in *35th Conference on Neural Information Processing Systems*, 2021, 1–24. doi: 10.48550/arXiv.2106.11189.

[29] Kull, M., Filho, T. de M. e S., & Flach, P., 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017. [Online]. Available: https://proceedings.mlr.press/v54/kull17a

[30] Motz, M., Krauß, J., & Schmitt, R. H., 2022. Benchmarking of hyperparameter optimization techniques for machine learning applications in production, *Advances in Industrial and Manufacturing Engineering*, Vol. 5, doi: 10.1016/j.aime.2022.100099.

[31] Murphy, A. H., 1973. A New Vector Partition of the Probability Score, *Journal of Applied Meteorology*, Vol. 12, No. 4, 595–600, doi: 10.1175/1520-

0450(1973)012<0595:ANVPOT>2.0.CO;2.

[32] Nasional, D. J. S., 2025. Monthly Report Monitoring JKN. [Online]. Available: https://kesehatan.djsn.go.id/kesehatan/doc/laporan-bulanan/Monthly_Report_JKN_06_2025.pdf

[33] Nembrini, S., Konig, I. R., & Wright, M. N., 2018. The revival of the Gini importance?, *Bioinformatics*, Vol. 34, No. 21, 3711–3718, doi: 10.1093/bioinformatics/bty373.

[34] Park, Y., & Ho, J. C., 2020. CaliForest: Calibrated Random Forest for Health Data, in *CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning*, Toronto, 2020, 40–50. doi: 10.1145/3368555.3384461.

[35] Powers, D. M. W., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, Vol. 2, No. 1, 37–63, doi: https://doi.org/10.48550/arXiv.2010.16061.

[36] Probst, P., Wright, M. N., & Boulesteix, A. L., 2019. Hyperparameters and tuning strategies for random forest, May 01, 2019, *Wiley-Blackwell*. doi: 10.1002/widm.1301.

[37] Rachael, M., 2020. Comparison of Elastic Net and Random Forest in identifying risk factors of stunting in children under five years of age in Kenya, University of Nairobi, 2020. [Online]. Available: https://erepository.uonbi.ac.ke/handle/11295/154085

[38] Saito, T., & Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE*, Vol. 10, No. 3, 1–21, doi: 10.1371/journal.pone.0118432.

[39] Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M., 2018. Comparison of Variable Selection Methods for Clinical Predictive Modeling, *International Journal of Medical Informatics*, Vol. 116, 10–17, doi: 10.1016/j.ijmedinf.2018.05.006.

[40] Schmidt, M., Le Roux, N., & Bach, F., 2017. Minimizing Finite Sums with the Stochastic Average Gradient, *Mathematical Programming*, Vol. 162, No. 1–2, 83–112, doi: 10.1007/s10107-016-1030-6.

[41] Seyam, E. A., 2025. Predicting High-Cost Healthcare Utilization Using Machine Learning : A Multi-Service Risk Stratification Analysis in EU-Based Private Group Health Insurance, *Risks*, Vol. 13, No. 7, 133, doi: 10.3390/risks13070133.

[42] Shrestha, A., Bergquist, S., Montz, E., & Rose, S., 2018. Mental Health Risk Adjustment with Clinical Categories and Machine Learning, *Health Services Research*, Vol. 53, 3189–3206, doi: 10.1111/1475-6773.12818.

[43] Si, Y., Sun, L., Chen, S., Fan, J., Pishgar, E., Alaei, K., Placencia, G., & Pishgar, M., 2025. Retrospective Machine Learning Approach for Forecasting In-Hospital Death in ICU Patients After Cardiac Arrest, *medRxiv*, 1–22, doi: 10.1101/2025.05.05.25327009.

[44] Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M., 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, *PLoS ONE*, Vol. 13, No. 8, 1–20, doi: 10.1371/journal.pone.0202344.

[45] Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., & Zeileis, A., 2008. Conditional variable importance for random forests, *BMC Bioinformatics*, Vol. 9, No. 307, 1–11, doi: 10.1186/1471-2105-9-307.

[46] Su, Y., Buist, D. S. M., Lee, J. M., Ichikawa, L., Miglioretti, D. L., Erin, J., Bowles, A., Wernli, K. J., Kerlikowske, K., Tosteson, A., Lowry, K. P., Henderson, L. M., Sprague, B. L., & Hubbard, R. A., 2023. Performance of statistical and machine learning risk prediction models for surveillance benefits and failures in breast cancer survivors, *Cancer Epidemiol Biomarkers Prev*, Vol. 32, No. 4, 561–571, doi: 10.1158/1055-9965.EPI-22-0677.

[47] Tay, J. K., Narasimhan, B., & Hastie, T., 2023. Elastic Net Regularization Paths for All Generalized Linear Models, *Journal of Statistical Software*, Vol. 106, No. 1, doi:

10.18637/jss.v106.i01.

[48] Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., & Chaimontree, S., 2021. Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand, *Informatics in Medicine Unlocked*, Vol. 26, doi: 10.1016/j.imu.2021.100769.

[49] Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society. Series B: Methodological*, Vol. 58, No. 1, 267–288, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[50] Timofeev, R., 2004. Classification and Regression Trees (CART) Theory and Applications, Humboldt University, Berlin, 2004.

[51] Varma, S., & Simon, R., 2006. Bias in error estimation when using cross-validation for model selection, *BMC Bioinformatics*, Vol. 7, No. 91, doi: 10.1186/1471-2105-7-91.

[52] Xiao, Y., Xing, E. P., & Neiswanger, W., 2022. Amortized Auto-Tuning: Cost-Efficient Bayesian Transfer Optimization for Hyperparameter Recommendation, *arXiv*, 1–23, doi: 10.48550/arXiv.2106.09179.

[53] Zela, A., Siems, J., Zimmer, L., Lukasik, J., Keuper, M., & Hutter, F., 2022. Surrogate NAS Benchmarks: Going Beyond the Limited Search Spaces of Tabular NAS Benchmarks, in *International Conference on Learning Representations*, 2022. doi: https://doi.org/10.48550/arXiv.2008.09777.

[54] Zhou, K., Hong, L., Hu, S., Zhou, F., Ru, B., Feng, J., & Li, Z., 2022. DHA: End-to-End Joint Optimization of Data Augmentation, *Transactions on Machine Learning Research*, 1–19, doi: 10.48550/arXiv.2109.05765.

[55] Zou, H., & Hastie, T., 2005. Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 67, No. 2, 301–320, [Online]. Available: http://www.jstor.org/stable/3647580