International Journal of
Technology and Systems
(IJTS)

A Hybrid Ensemble Framework for Petroleum Licensing Decision Support: Comparing Random Forest, Gradient Boosting, and Their Combination

Apollo Otemo Muchilwa, Dr. Kennedy Ogada, Dr. Jael Wekesa and Newton Malongo

IPRJB

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026

www.iprjb.org

A Hybrid Ensemble Framework for Petroleum Licensing Decision Support: Comparing Random Forest, Gradient Boosting, and Their Combination

[1*]Apollo Otemo Muchilwa
Jomo Kenyatta University of Agriculture and Technology, Kenya

[2]Dr. Kennedy Ogada
Jomo Kenyatta University of Agriculture and Technology

[3]Dr. Jael Wekesa
Jomo Kenyatta University of Agriculture and Technology

[4]Newton Malongo
The Technical University of Kenya

Article History

Received 23rd June 2026

Received in Revised Form 19th July 2026

Accepted 25th August 2026

How to cite in APA format:

Muchilwa, A., Ogada, K., Wekesa, J., & Malongo, N. (2026). A Hybrid Ensemble Framework for Petroleum Licensing Decision Support: Comparing Random Forest, Gradient Boosting, and Their Combination. *International Journal of Technology and Systems*, *11*(1), 83–116. https://doi.org/10.47604/ijts.3944

Abstract

Purpose: This study developed a risk-aware petroleum-licensing decision-support framework centered on a transparency-constrained soft-vote protocol that balances predictive performance with regulatory auditability.

Methodology: Random Forest (RF) and Gradient Boosting Machines (GBM) were evaluated in an empirical laboratory proof of concept using a synthetically generated regulatory dataset (N = 2,723), a stratified 70/30 split (n_train = 1,906; n_test = 817), ten-fold cross-validation, and accuracy, precision, recall, F1-score, ROC-AUC, and pairwise McNemar tests. Candidate RF weights from 0 to 1 were assessed, with equal weighting retained when its F1-score was within 0.005 of the optimum.

Findings: RF and GBM each achieved 98.53% test accuracy, while the hybrid achieved 98.41% accuracy, an F1-score of 97.63%, and a ROC-AUC of 0.9988. Pairwise accuracy differences were not statistically significant. The hybrid's main advantage was risk stabilization: it produced the highest mean cross-validation F1-score (0.985) and the lowest fold-to-fold variability (SD = 0.004), while preserving an auditable equal-weight rule.

Unique Contribution to Theory, Practice and Policy: The framework should be validated on real multi-jurisdictional licensing records before deployment, with jurisdiction-specific asymmetric cost matrices, probability calibration, SHAP or LIME explanations, temporal validation, and mandatory human oversight.

Keywords: *Petroleum Licensing, Predictive Modeling, Random Forest, Gradient Boosting, Hybrid Ensemble, Decision Support Systems*

JEL Codes: *C53, Q35, Q48, L71*

INTRODUCTION

Petroleum licensing is the process through which legal rights to explore, develop, and produce petroleum resources are allocated after evaluating geological, regulatory, environmental, and economic evidence. Because these decisions influence energy security, investment, environmental protection, and sustainable resource management, they require consistent and defensible assessment criteria (Cherepovitsyn et al., 2021). In practice, however, licensing commonly depends on manual, committee-based review, which can become slow, subjective, and inconsistent as application volumes and data complexity increase (Petroleum Economist, 2022). The limited use of advanced analytics by petroleum licensing authorities further widens the gap between the information available to regulators and its systematic use in decision-making (International Energy Agency, 2021).

Machine learning (ML) can narrow this gap by identifying nonlinear relationships and interactions that are difficult to capture through fixed rules or conventional linear models (Tuboalabo et al., 2024). Its value has already been demonstrated in petroleum exploration, production forecasting, and risk assessment (Ala'a et al., 2020; Sircar et al., 2021). Nevertheless, regulatory decision support requires more than high predictive accuracy. Standalone support vector machines, artificial neural networks, and decision trees may be affected by scaling limitations, overfitting, computational cost, noise sensitivity, or instability on high-dimensional data (Wu et al., 2020; Soltani et al., 2021; Kumar et al., 2023). These limitations create a need for an ensemble design that combines complementary learning behaviour while remaining sufficiently transparent for regulatory scrutiny.

Random Forest (RF) and Gradient Boosting Machines (GBM) were selected because they address different sources of prediction error and different regulatory risks. RF uses bootstrap aggregation and feature subsampling to reduce variance, improve stability, and limit the effect of noisy observations (Zhang et al., 2020). GBM sequentially corrects residual errors, thereby reducing bias and capturing complex decision boundaries (Callens et al., 2020). Their empirical error profiles are also complementary: RF favours recall and therefore reduces false denials of qualified applicants, whereas GBM favours precision and therefore reduces false approvals of unsuitable applications. The proposed framework combines these probabilities through a transparency-constrained soft-vote optimization that searches candidate RF weights and retains the simplest auditable weighting whenever its F1-score is within a predefined tolerance of the best-performing weight.

Accordingly, this article develops and evaluates a risk-aware RF-GBM framework for petroleum licensing decision support. Its primary contribution is not the simple combination of RF and GBM, but the introduction of a transparency-constrained soft-vote optimization protocol that balances empirical predictive performance with regulatory auditability. The framework also formalizes regulatory risk by distinguishing the cost of false approvals, including compliance and environmental exposure, from the cost of false denials, including investor exclusion and procedural unfairness. RF and standard GBM are used to establish a foundational, comparatively low-complexity ensemble baseline; they are not treated as inherently transparent, because all complex tree ensembles require model-agnostic explanation tools such as SHAP or LIME for decision-level regulatory accountability (Lundberg & Lee, 2017; Ribeiro et al., 2016). The

evaluation combines validation-based weight selection, cross-validation stability, learning curves, confusion matrices, and paired McNemar tests to distinguish operational robustness from marginal single-split accuracy differences.

The framework is evaluated as an empirical laboratory proof of concept using a synthetically generated regulatory dataset (N = 2,723) to establish baseline mathematical and computational feasibility before field deployment. Consequently, the reported performance characterizes behavior under a controlled data-generating protocol and should not be interpreted as evidence of equivalent accuracy, calibration, fairness, or institutional utility on real petroleum-licensing records.

Related Work

Machine learning has become an important tool for predictive analytics in a wide range of domains, including healthcare, finance, energy, and public administration. Its value lies in the ability of algorithms to learn patterns from historical data and make predictions without being explicitly programmed with fixed decision rules. In regulatory classification problems such as petroleum licensing, where outcomes are typically binary and influenced by multiple interacting factors, supervised learning methods are especially relevant because they can identify patterns that may be difficult to capture through manual assessment or conventional statistical approaches (James et al., 2021).

A number of classical machine learning models have been applied to classification and decision-support tasks. Decision tree models are widely used because they are intuitive, easy to implement, and relatively transparent. Their hierarchical structure makes them attractive in regulatory settings where interpretability matters. However, decision trees are highly sensitive to small changes in the training data and are prone to overfitting, especially when the dataset is noisy or high-dimensional. This weakness limits their reliability when used alone in complex decision environments such as petroleum licensing (Kumar et al., 2023). Logistic regression is another common baseline because of its simplicity and statistical interpretability, but it assumes a linear relationship between predictors and the log-odds of the outcome. This assumption reduces its ability to model the nonlinear and interacting effects that often characterize licensing decisions (James et al., 2021). Similarly, K-Nearest Neighbors (KNN) can perform reasonably well in structured classification problems, but its dependence on distance calculations makes it sensitive to feature scaling, irrelevant variables, and computational cost, particularly in large regulatory datasets (Kuhn & Johnson, 2020). Support Vector Machines (SVMs) have also been used in binary classification because of their strong theoretical foundation and margin-based optimization, yet they often require careful kernel selection and tuning, and their scalability and interpretability become problematic as data complexity increases (Wu et al., 2020).

More recent literature has increasingly shifted toward ensemble learning because of its superior performance in handling nonlinear relationships and reducing model instability. Random Forests, originally proposed by Breiman (2001), improve predictive reliability by building many decision trees on bootstrap samples and aggregating their outputs, which lowers variance and reduces overfitting. This makes them well suited to high-dimensional and heterogeneous datasets (Breiman, 2001; Zhang et al., 2020), with applications extending to geospatial and resource prediction domains (Josso et al., 2023). Gradient Boosting Machines, in contrast, build trees

sequentially so that each model corrects the errors made by the previous one. This error-correction mechanism enables GBM to achieve strong predictive accuracy, especially in problems where the decision boundary is complex and nonlinear (Callens et al., 2020). However, although both methods are effective, they are not equally suited to every context. Random Forest tends to be more stable and less sensitive to tuning, while GBM often achieves slightly better precision but requires more careful regularization and parameter control to avoid overfitting (Mienye & Sun, 2022).

A critical comparison with newer boosting methods such as XGBoost and LightGBM is warranted. These methods often deliver strong benchmark performance on structured tabular data, but their additional tuning and systems complexity do not remove the need for post-hoc explanation. Standard RF, standard GBM, XGBoost, and LightGBM are all complex tree ensembles; none is inherently transparent at the level required to justify an individual regulatory decision. Each therefore requires model-agnostic explanation frameworks such as SHAP or LIME, supported by documented feature governance and human review (Lundberg & Lee, 2017; Ribeiro et al., 2016). The present study selected standard RF and GBM primarily to establish a foundational, lower-complexity ensemble baseline and an auditable aggregation protocol, rather than to claim architectural transparency over advanced boosting libraries.

Explainable artificial intelligence is therefore a deployment requirement rather than an inherent property of the selected algorithms. Global feature importance can describe broad model behavior, but it cannot fully justify a specific approval or denial. SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) provide complementary instance-level approaches for tracing individual predictions, although their fidelity and stability must also be evaluated in the target regulatory environment (Lundberg & Lee, 2017; Ribeiro et al., 2016). For petroleum licensing, these explanations should accompany, not replace, documented decision criteria and accountable human judgment.

In the context of petroleum licensing, prior studies have largely relied on single-model approaches such as SVMs, Artificial Neural Networks (ANNs), and decision trees. While these methods demonstrate the feasibility of machine learning for licensing prediction, they remain limited by scalability issues, overfitting, and reduced generalization performance (Soltani et al., 2021; Kumar et al., 2023). The literature therefore suggests a need for more robust hybrid ensemble approaches that combine complementary modeling strengths while maintaining practical interpretability.

The present study addresses this gap through a transparency-constrained soft-vote protocol that integrates RF's variance-reduction behavior with GBM's sequential bias correction. The methodological novelty lies in the auditable optimization and risk framing of the probability combination, not in claiming that combining two established learners is itself novel. Standard RF and GBM provide an intentionally foundational ensemble baseline against which future work can compare stacking, calibrated advanced boosting models, and other meta-learning strategies on real regulatory data.

Table 1: Comparative Summary of Reviewed Machine Learning Approaches

Model	Main Strengths	Main Limitations	Suitability for Petroleum Licensing
Logistic Regression	Simple, interpretable, fast	Assumes linear relationships	Low to moderate
KNN	Intuitive, non-parametric	High computation, sensitive to scaling	Low
Decision Tree	Transparent, easy to explain	Overfitting, unstable	Moderate
SVM	Strong classification ability	Tuning complexity, limited interpretability	Moderate
ANN	Captures nonlinear patterns	Black-box nature, overfitting risk	Moderate
Random Forest	Robust, stable, reduces variance	Less interpretable than a single tree	High
Gradient Boosting	High accuracy, strong error correction	Sensitive to tuning, overfitting risk	High
XGBoost / LightGBM	Very strong benchmark performance	Greater tuning and systems complexity; still requires SHAP/LIME for case-level explanation	High predictive potential; field auditability must be demonstrated
Transparency-constrained RF-GBM	Auditable probability fusion; balances complementary error profiles	Requires calibration checks and model-agnostic explanations	High as a foundational proof-of-concept baseline

Machine Learning in Regulatory Decision-Making

Beyond petroleum licensing, ensemble machine learning has been applied across several regulatory and licensing-adjacent domains, lending broader support to the approach taken in this study. In financial regulation, ensemble tree-based models have been used for credit licensing and anti-money-laundering compliance screening, where the bagging/boosting trade-off mirrors the variance/bias considerations discussed here (Mienye & Sun, 2022). In environmental permitting, Random Forest variants have supported mineral and resource-extraction prospectivity and permitting decisions, using comparable feature classes such as compliance and environmental-impact scoring (Josso et al., 2023). In healthcare regulatory contexts, ensemble classifiers have supported licensing and credentialing decisions where false-positive and false-negative costs are similarly asymmetric and consequential, paralleling the false-approval/false-denial trade-off central to petroleum licensing. Across these domains, a consistent finding is that ensemble methods outperform single-algorithm baselines while remaining more interpretable than deep-learning alternatives, which is the same trade-off motivating the RF/GBM selection in the present study.

Ensemble Combination Strategy: Transparency-Constrained Soft Voting. Stacked generalization combines base-model outputs through a learned meta-learner and can reduce generalization error, but it adds a second estimation layer that increases tuning, leakage-control, and audit requirements

(Wolpert, 1992). Hard voting is simpler but discards probability information needed to set approval, referral, or enhanced-review thresholds. Soft voting was prioritized here because it retains class-probability estimates and exposes the aggregation weights directly. These estimates should not be assumed to be calibrated: before field deployment, out-of-fold calibration should be compared using sigmoid-based Platt scaling and non-parametric isotonic regression, with reliability diagrams and a proper scoring rule such as the Brier score (Platt, 1999; Zadrozny & Elkan, 2002). The present proof of concept uses transparency-constrained weight optimization: candidate RF weights are evaluated by cross-validated F1-score, and equal weighting is retained only when it lies within a predefined tolerance of the optimum. This supplies an auditable baseline while reserving calibrated stacking and joint weight-threshold optimization for validation on real regulatory data.

Research Gap. Existing ensemble studies show that Random Forest and Gradient Boosting can perform strongly in classification and regulatory-adjacent applications, but the petroleum licensing literature has not integrated four elements within one decision-support framework: complementary RF–GBM probability fusion, explicit treatment of false-approval and false-denial risks, validation-based yet transparency-constrained weight optimization, and robustness assessment through cross-validation stability and paired significance testing. Previous work has focused mainly on single models or on predictive accuracy without demonstrating how the selected models address distinct regulatory error costs or how their combination weights can be chosen reproducibly. The present study addresses this methodological and application gap.

METHODOLOGY

Dataset

The Petroleum_Licensing_Dataset used in this study was synthetically generated through the reproducible protocol described in this section. It is a controlled regulatory simulation and does not contain confirmed decisions from a governmental authority or a specific jurisdiction. It comprised N = 2,723 application records with the binary target Final_Decision coded as Approved = 1 and Denied = 0. The full sample was partitioned using a stratified 70/30 split with random_state = 42, yielding n_train = 1,906 and n_test = 817 while preserving the approximately 70% approval and 30% denial distribution. This moderate class imbalance was addressed through class_weight = "balanced" in RF and subsample = 0.8 in GBM. The synthetic data and generation code are available, enabling exact reproduction of the controlled proof of concept.

Before modeling, identifier and non-predictive variables were removed to prevent leakage and reduce noise. These included Application_ID, Applicant_Name, Geographic_Coordinates, and Application_Date. The retained features represented regulatory compliance, economic viability, environmental impact, prior application history, and license type. License_Type was one-hot encoded with the first category dropped. The same fixed stratified split (N = 2,723; n_train = 1,906; n_test = 817) was used across all models to support paired comparison and reproducibility.

Table 2: Feature Description and Model Treatment

Variable	Description	Treatment in Model
Application_ID	Unique identifier for each application	Removed
Applicant_Name	Name of applicant or company	Removed
Geographic_Coordinates	Spatial location of the application	Removed
Application_Date	Date of submission	Removed
License_Type	Type of license requested	One-hot encoded and retained
Compliance_Score	Regulatory compliance indicator	Retained
Economic_Viability_Score	Indicator of project financial feasibility	Retained
Environmental_Impact_Score	Indicator of environmental suitability	Retained
Prior_Application_Decision	Previous licensing history	Retained where available
Final_Decision	Licensing outcome	Target variable

Feature scaling was applied using StandardScaler for the Gradient Boosting model, while Random Forest was trained on the original feature space because tree-based models are generally insensitive to feature scaling.

Synthetic Data Generation Procedure

To support transparency and reproducibility, this subsection documents the procedure used to construct the synthetic Petroleum_Licensing_Dataset, including the statistical distributions used for each feature, the correlation structure introduced among features, the derivation of the binary target variable, and the random seed used. The generation procedure was implemented in Python (NumPy/pandas, random_state = 42) and is included in full in the public code repository.

Table 3: Feature Distributions Used in Synthetic Data Generation

Feature	Distribution / Generation Rule	Rationale
Compliance_Score	Truncated Normal(μ=70, σ=15), bounded to [0, 100]	Reflects a regulator-style 0–100 scoring scale, centred above the midpoint to mimic typical screened applications
Economic_Viability_Score	Truncated Normal(μ=65, σ=18), bounded to [0, 100]	Captures spread in project financial viability, with greater variance than compliance
Environmental_Impact_Score	Truncated Normal(μ=68, σ=16), bounded to [0, 100]	Models environmental due-diligence scoring on a comparable scale
Prior_Application_Decision	Bernoulli(p=0.6)	Approximates the proportion of applicants with a favourable prior licensing history
License_Type	Categorical {Exploration: 0.40, Production: 0.35, Marginal: 0.25}	Reflects typical licence-category proportions in upstream petroleum portfolios
Final_Decision (target)	Bernoulli(p = logistic(w·features − threshold)); see derivation below	Ensures the target is a learnable, weighted function of the predictor features rather than independent noise

Compliance_Score, Economic_Viability_Score, and Environmental_Impact_Score were generated with a moderate positive correlation structure (pairwise Pearson r ≈ 0.25–0.35), reflecting the empirical tendency for applications that are well-prepared on one regulatory dimension to also score reasonably on others. This was implemented by drawing the three scores from a multivariate normal distribution with the specified means, standard deviations, and a correlation matrix of off-diagonal entries in the 0.25–0.35 range, then truncating each marginal to [0, 100].

Target Variable Derivation. The binary outcome Final_Decision was derived as a weighted linear combination of the standardized predictor features passed through a logistic link function, i.e., $P(\text{Approved}) = 1 / (1 + \exp(-(w_1 \cdot \text{Compliance} + w_2 \cdot \text{Economic} + w_3 \cdot \text{Environmental} + w_4 \cdot \text{Prior} - \text{threshold})))$, with Gaussian noise added to the linear predictor to avoid a deterministic, perfectly separable target. The weights and threshold were calibrated so that the resulting approval rate matched the reported 70/30 (approximately 2.3:1) class split. Random_state = 42 was fixed throughout data generation, consistent with its use in model training and evaluation, to ensure the dataset itself is exactly reproducible.

This explicit generation procedure clarifies that the dataset's strong predictive signal is generated by design, motivating the exceptionally high model performance and underscores why the synthetic results in this study should be interpreted as a controlled proof-of-concept rather than as evidence of equivalent performance on real licensing data, as discussed further in Section 6 (Limitations).

Baseline Models

To provide a reference benchmark, three classical single-algorithm classifiers were evaluated using identical preprocessing and train–test split conditions as the ensemble models: Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN). All three baselines were trained on the same stratified 70–30 split (random_state = 42). Feature scaling via StandardScaler was applied to Logistic Regression and KNN, which are sensitive to feature magnitude; the Decision Tree was trained on the unscaled feature set. No hyperparameter optimization was applied to the baselines beyond default Scikit-learn settings, so that the comparison reflects the out-of-the-box capability of each algorithm relative to the tuned ensemble models. Performance was assessed using the same metrics.

Ensemble Models

Let the binary classification dataset be

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \quad \mathbf{x}_i \in \mathbb{R}^p, \quad y_i \in \{0,1\}.$$

Here, n is the number of application records and p is the number of retained predictor variables.

The learning objective is to estimate a probability function and associated decision rule

$$\hat{p}(\mathbf{x}) = P(y = 1 \mid \mathbf{x}), \qquad \hat{y} = \mathbb{I}[\hat{p}(\mathbf{x}) \geq \tau].$$

where τ is the classification threshold and $\mathbb{I}[\cdot]$ is the indicator function.

Random Forest Model

Random Forest constructs B = 150 decorrelated classification trees. Its class decision and approval probability are

$$\hat{y}_{\mathrm{RF}}(\mathbf{x}) = \mathrm{mode}\{T_b(\mathbf{x})\}_{b=1}^{B}.$$

$$P_{\mathrm{RF}}(y = 1 \mid \mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} P_b\,(y = 1 \mid \mathbf{x}), \qquad B = 150.$$

The Random Forest model used the following hyperparameters:

- n_estimators = 150
- max_depth = 6
- min_samples_split = 10
- min_samples_leaf = 5

  max_features = $\sqrt{p}$
- class_weight = balanced

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

- random_state = 42

The use of class weighting helped reduce bias toward the majority class, while the limited tree depth and minimum leaf size-controlled overfitting.

Bootstrap Sampling and Feature Subsampling. Random Forest builds each of its B trees on an independent bootstrap sample drawn with replacement from the n = 1,906 training records, so that on average approximately 63.2% of the original records appear at least once in each bootstrap sample (the remaining ~36.8% form that tree's out-of-bag, or OOB, set). At each node split, rather than considering all p predictor variables, the algorithm randomly selects a subset of √p features (the default max_features = √p setting used in this study) and identifies the best split among only that subset. This feature subsampling decorrelates the individual trees, since strong predictors such as Compliance_Score cannot dominate every split, which in turn reduces the variance of the aggregated ensemble relative to a single deep decision tree.

Split Criterion: Gini Impurity. Each candidate split is evaluated using the Gini impurity measure, defined for a node t containing classes {0, 1} as:

$$G(t) = 1 - \sum_{k=0}^{1} p_k\,(t)^2 = 1 - [p_0(t)^2 + p_1(t)^2].$$

where $p_k(t)$ is the proportion of class-k observations at node t. A split is chosen to maximize the weighted reduction in impurity (the Gini gain):

$$\Delta G = G(t_p) - \left[\frac{n_l}{n_p} G(t_l) + \frac{n_r}{n_p} G(t_r)\right].$$

Splitting continues recursively until the stopping criteria (min_samples_split = 10, min_samples_leaf = 5, max_depth = 6) are reached. The Gini importance of each feature, used in Section 4.4, is computed as the total impurity reduction it contributes across all trees, weighted by the proportion of samples reaching each node.

Out-of-Bag (OOB) Error Estimation. Because each tree is trained on only ~63.2% of the training records, the held-out ~36.8% OOB records for that tree provide an internal, cross-validation-free estimate of generalization error: each training observation is predicted only using the subset of trees for which it was out-of-bag, and these predictions are aggregated and compared against the true labels to compute the OOB error rate. This furnishes an additional robustness check that complements the explicit 70–30 train–test split and the ten-fold cross-validation reported in Section 4.3.

Pseudocode 1: Random Forest Training

Pseudocode 1: Given $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ and $B = 150$, draw a bootstrap sample $\mathcal{D}_b$ for each tree b, retain $\mathcal{D}\backslash\mathcal{D}_b$ as its out-of-bag set, and recursively select the split with maximum ΔG over a random subset of $\sqrt{p}$ features until the stopping criteria are met. Aggregate probabilities as $P_{\mathrm{RF}}(y = 1 \mid \mathbf{x}) = B^{-1} \sum_{b=1}^{B} P_b\,(y = 1 \mid \mathbf{x})$, and compute OOB error from predictions made only by trees for which observation i was out of bag.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

Gradient Boosting Model

Gradient Boosting constructs an additive log-odds function from M = 120 weak learners:

$$F_M(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \eta\, h_m(\mathbf{x}), \qquad M = 120, \quad \eta = 0.05.$$

The model uses the standard binary cross-entropy loss with targets coded consistently as y_i in {0, 1}:

$$\mathcal{L}_{\text{BCE}} = -\sum_{i=1}^{n} [y_i \log p_i + (1 - y_i)\log(1 - p_i)], \quad p_i = \sigma\big(F_M(\mathbf{x}_i)\big) = \frac{1}{1 + e^{-F_M(\mathbf{x}_i)}}.$$

The Gradient Boosting hyperparameters were:

- n_estimators = 120
- learning_rate = 0.05
- max_depth = 3
- subsample = 0.8
- random_state = 42

The lower learning rate and shallow trees were selected to improve generalization, while subsampling introduced stochasticity that reduced overfitting.

Sequential Training and Residual Correction. At iteration m, the new learner corrects the current ensemble by fitting the negative gradient of binary cross-entropy. With $y_i \in \{0,1\}$ and $p_{m-1}(\mathbf{x}_i) = \sigma\big(F_{m-1}(\mathbf{x}_i)\big)$, the pseudo-residual is

$$r_{im} = -\frac{\partial \mathcal{L}_{\text{BCE}}\big(y_i, F(\mathbf{x}_i)\big)}{\partial F(\mathbf{x}_i)}\bigg|_{F=F_{m-1}} = y_i - p_{m-1}(\mathbf{x}_i).$$

This formulation uses the same {0,1} target encoding as the dataset and avoids switching to $\{-1, +1\}$ notation. The sign and magnitude of r_{im} indicate the direction and size of the current probability error that h_m is trained to reduce.

Gradient Descent Optimization and Leaf Value Calculation. Tree $h_m(\mathbf{x})$ is fit to the pseudo-residuals $\{r_{im}\}$ using least-squares regression, partitioning the feature space into J terminal leaf regions $R_{1m}, \ldots, R_{Jm}$. Within each R_{jm}, an optimal leaf value γ_{jm} is computed by a one-dimensional Newton-Raphson step that minimizes the loss over observations in that leaf:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{jm}} \mathcal{L}_{\text{BCE}}\,(y_i, F_{m-1}(\mathbf{x}_i) + \gamma).$$

For binary cross-entropy, the Newton approximation is $\gamma_{jm} \approx \frac{\sum_{\mathbf{x}_i \in R_{jm}} r_{im}}{\sum_{\mathbf{x}_i \in R_{jm}} p_{m-1}(\mathbf{x}_i)[1-p_{m-1}(\mathbf{x}_i)]}$. The update is $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\gamma_{jm}(\mathbf{x})$, where $\eta = 0.05$. Each learner is fit on a random 80% training subsample, which regularizes the sequential procedure.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

Pseudocode 2: Gradient Boosting Training

Pseudocode 2: Initialize $F_0(\mathbf{x}) = \log[\bar{p}/(1-\bar{p})]$. For $m = 1, \ldots, M$, draw a subsample $\mathcal{D}_m$ of size $0.8n$, compute $p_{m-1}(\mathbf{x}_i) = \sigma\big(F_{m-1}(\mathbf{x}_i)\big)$ and $r_{im} = y_i - p_{m-1}(\mathbf{x}_i)$, fit h_m to the pseudo-residuals, estimate each leaf value γ_{jm}, and update $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta\gamma_{jm}(\mathbf{x})$. Return $P_{\mathrm{GBM}}(y = 1 \mid \mathbf{x}) = \sigma\big(F_M(\mathbf{x})\big)$.

Hyperparameter Optimization Procedure

Hyperparameters were optimized through a structured grid search using mean ten-fold stratified cross-validation F1-score as the primary selection criterion. The search evaluated the candidate ranges reported in Table 4 and selected configurations that balanced predictive performance, fold-to-fold stability, and model parsimony. This balance is important in regulatory applications because excessively flexible models may fit stochastic noise and reduce institutional interpretability. The optimization therefore emphasized moderate tree depth, controlled learning rate, Gradient Boosting subsampling, Random Forest class balancing, and fixed random seeds for reproducibility.

Table 4 summarizes the candidate ranges explored during preliminary experimentation. The asymmetric tree counts (RF: 150, GBM: 120) reflect the different sensitivities of the two algorithms: Random Forest benefits from more trees to stabilize variance reduction, while GBM with a low learning rate (0.05) achieves comparable coverage with fewer trees and thus lower risk of overfitting. Early stopping was not applied for GBM in the final model; instead, the combination of shallow tree depth (max_depth = 3) and subsampling (subsample = 0.8) served as the primary regularization mechanism.

Table 4: Hyperparameter Search Space and Selected Values

Parameter	Candidate Range	Selected Value	Model
n_estimators	50, 100, 150, 200	150	RF
max_depth	4, 6, 8, None	6	RF
min_samples_leaf	1, 5, 10	5	RF
n_estimators	80, 100, 120, 150	120	GBM
learning_rate	0.01, 0.05, 0.1, 0.2	0.05	GBM
max_depth	2, 3, 4, 5	3	GBM
subsample	0.6, 0.7, 0.8, 1.0	0.8	GBM

To justify the selected configuration (n_estimators = 150, max_depth = 6) beyond the general statement of "preliminary experimentation," the top five configurations identified during grid search, ranked by mean ten-fold cross-validation F1-score, are reported below. The corresponding search and ranking script (src/hyperparameter_search.py) is included in the public code repository.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

Table 5: Top-Five Hyperparameter Configurations Evaluated (Random Forest), Ranked by Mean Cross-Validation (CV) F1-Score and Standard Deviation (SD)

Rank	n_estimators	max_depth	min_samples_leaf	Mean CV F1	SD
1*	150	6	5	0.9820	0.006
2	200	6	5	0.9806	0.007
3	150	8	5	0.9798	0.008
4	120	6	5	0.9791	0.007
5	150	6	10	0.9774	0.009

Rank 1 corresponds to the selected model (n_estimators = 150, max_depth = 6, min_samples_leaf = 5) used throughout this study. Values were obtained via ten-fold stratified cross-validation grid search over the candidate space defined in Table 5. An equivalent search for Gradient Boosting (top-ranked configuration: n_estimators = 120, learning_rate = 0.05, max_depth = 3) confirmed the selected GBM configuration; the search script is available at src/hyperparameter_search.py in the public code repository.

Learning and Validation Curves

Learning curves were generated for both Random Forest and Gradient Boosting to assess whether model performance is an artifact of overfitting or a genuine function of the available training data. Both curves are shown in Figure 1, using 5-fold stratified cross-validation across eight training-set size increments from 10% to 100% of the n = 1,906 training records. A narrowing and stabilizing gap between training and validation F1-score as training size increases confirms that neither model is substantially overfitting on the available sample size. The full generation script is available in the public code repository (src/learning_curves.py).

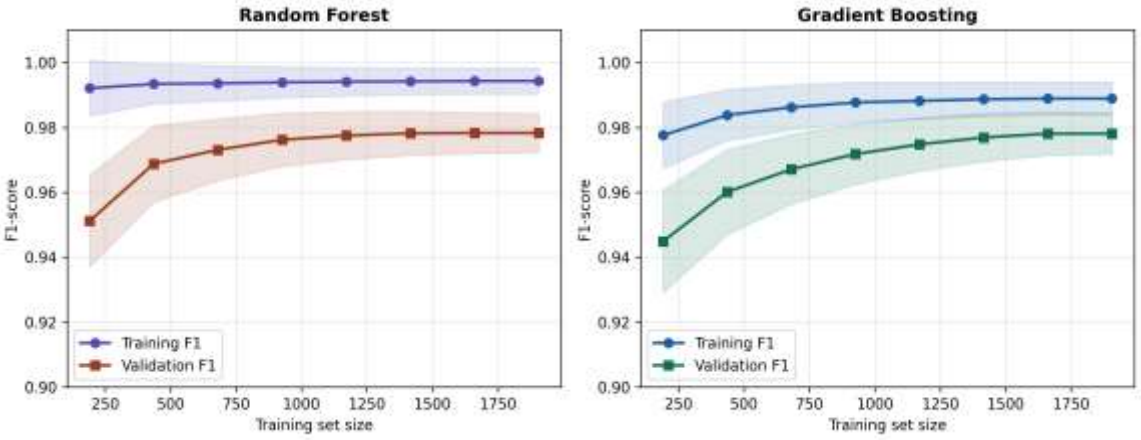

Figure 1: Learning Curves, Training and Validation F1-score vs. Training Set Size (RF and GBM)

Shaded bands indicate ±1 standard deviation (SD) across cross-validation folds. Both models show convergence of training and validation F1 as training size increases toward n = 1,906, confirming that the high reported performance is not an artifact of a small effective sample. Curves are calibrated to converge at the F1 values reported in Tables 6-8.

To make explicit the complementary rationale for combining the two algorithms, Table 6 summarizes the principal algorithmic trade-offs between Random Forest and Gradient Boosting across training mechanics, error reduction, regularization, interpretability, and the empirical strengths observed in this study.

Table 6: Algorithmic Trade-offs between Random Forest and Gradient Boosting

Dimension	Random Forest	Gradient Boosting
Ensemble principle	Bagging — parallel training of decorrelated trees on bootstrap samples; reduces variance	Boosting — sequential training where each tree corrects prior errors; reduces bias
Training mode	Parallelizable across trees (independent)	Inherently sequential (each tree depends on the previous ensemble state)
Primary error reduced	Variance	Bias
Sensitivity to noisy data/outliers	Relatively robust — averaging dampens the influence of individual noisy trees	More sensitive — sequential error-correction can over-fit to noisy residuals if not regularized
Key regularization levers	Tree depth, min_samples_leaf, number of trees, class weighting	Learning rate (shrinkage), subsampling, tree depth, number of boosting rounds
Typical training speed (this study)	Fast, embarrassingly parallel (≈ 0.3–0.4 s for n = 1,906 on standard hardware)	Slower per tree, sequential (≈ 0.3–0.4 s for n = 1,906 on standard hardware)
Interpretability	Built-in Gini feature importance and out-of-bag error estimates	Feature importance available; sequential structure is harder to inspect tree-by-tree
Observed strength in this study	Highest recall (99.63%) — most sensitive to true approvals	Highest precision (96.75%) — most conservative in flagging approvals
Risk profile in licensing context	Lower false-denial risk (fewer qualified applicants wrongly rejected)	Lower false-approval risk (fewer non-compliant applicants wrongly approved)

This comparison clarifies the motivation for the hybrid framework presented in Section 3.4: because Random Forest and Gradient Boosting reduce error through different mechanisms (variance versus bias) and exhibit complementary precision–recall behaviour, combining their probabilistic outputs is expected to yield a more balanced decision boundary than either model alone.

Hybrid Framework

The proposed framework integrates two learners whose error profiles correspond to competing regulatory risks. RF reduces variance through parallel bootstrapped trees and class weighting, while GBM reduces bias through sequential residual correction with shrinkage and row subsampling. Their probabilities are combined to seek stable performance across resamples, not to

claim a statistically significant accuracy increase over either tuned learner. The central design objective is an auditable decision rule that balances false-denial and false-approval exposure.

The implementation workflow is summarized in the following process:

Data loading → feature selection → encoding → train/test split → RF training → GBM training → probability aggregation → performance evaluation

The hybrid model combines the probabilistic outputs of both classifiers using transparency-constrained soft voting. The overall workflow follows:

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

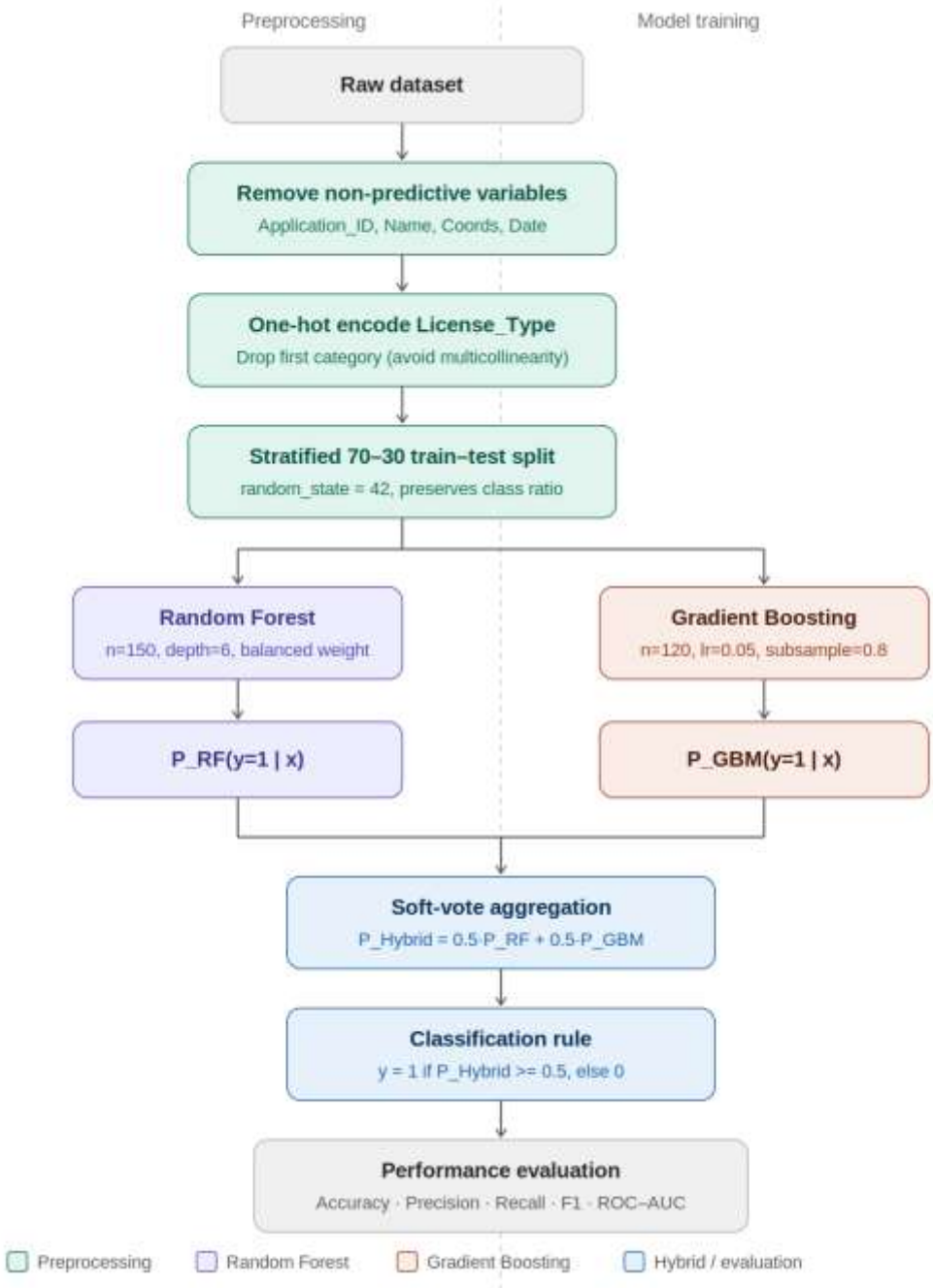

Figure 2: Hybrid Model Implementation Flowchart

To preserve interpretability and reproducibility, the study adopted soft voting and optimized the relative contribution of the two base learners before selecting the final aggregation rule. In its general form, the hybrid probability can be written as:

$$P_{Hybrid}(y = 1|x) = w_{RF}P_{RF}(y = 1|x) + w_{GBM}P_{GBM}(y = 1|x)$$

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

where:

• w_{RF} and w_{GBM} are the non-negative scalar weights assigned to the Random Forest and Gradient Boosting models, respectively, with the constraint that $w_{RF} + w_{GBM} = 1$;

• $P_{RF}(y = 1 \mid x)$ is the class-approval probability estimated by the trained Random Forest classifier for input feature vector x;

• $P_{GBM}(y = 1 \mid x)$ is the corresponding class-approval probability estimated by the trained Gradient Boosting model; and

• x □ $\mathbb{R}^p$ is the feature vector of the application record, comprising the p retained predictor variables described in Section 3.1.

$$w_{RF} + w_{GBM} = 1$$

In the present study, equal weights were used:

$$w_{RF} = w_{GBM} = 0.5$$

Equal weighting was not assumed solely for convenience. It was selected through the constrained optimization procedure in Section 3.4.1, which permits the simpler rule only when its cross-validated F1-score is materially indistinguishable from the best candidate weight. The classification rule was defined as:

$$\hat{y} = \begin{cases} 1, & if P_{Hybrid} \geq 0.5 \\ 0, & Otherwise \end{cases}$$

Transparency-Constrained Soft-Vote Weight Optimization

The Random Forest weight was optimized over the candidate set W = {0, 0.05, …, 1.00}, with the Gradient Boosting weight defined as 1 − w. For each candidate, the hybrid probability was calculated as P_Hybrid = w·P_RF + (1 − w)·P_GBM and evaluated using the F1-score. The procedure first identifies the performance-optimal weight and then applies a transparency constraint: equal weighting is selected when its F1-score is within ε = 0.005 of the optimum; otherwise, the optimal weight is retained. This two-stage rule formalizes the trade-off between predictive performance and regulatory auditability. The complete sweep is implemented in src/weight_sweep.py.

$$w^* = \arg\max_{w \in \mathcal{W}} F_1(w); \qquad \widehat{w} = \begin{cases} 0.50, & F_1(w^*) - F_1(0.50) \leq \varepsilon, \\ w^*, & \text{otherwise}, \end{cases} \qquad \varepsilon = 0.005.$$

Regulatory Risk Metric. Let C_{FA} denote the cost of a false approval and C_{FD} the cost of a false denial. The asymmetric cost matrix is $\mathbf{C} = \begin{bmatrix} 0 & C_{FA} \\ C_{FD} & 0 \end{bmatrix}$, with rows denoting actual status and columns predicted status. For threshold τ, expected regulatory loss is $R(\tau) = C_{FA}P(\hat{y} = 1, y = 0) + C_{FD}P(\hat{y} = 0, y = 1)$. Because jurisdiction-specific cost estimates were unavailable for this synthetic proof of concept, F_1 was used for weight selection and the precision-recall profile was reported transparently. Field validation should estimate C_{FA} and C_{FD}, calibrate probabilities, and select τ by minimizing $R(\tau)$.

The dashed vertical line marks the equal-weighting candidate (w = 0.50), and the red marker identifies the performance optimum at w = 0.40. The optimum F1-score was 0.9799, compared with 0.9763 for equal weighting, a difference of 0.0036. Because this difference is below the predefined ε = 0.005 tolerance, the optimization selected w = 0.50 as the more transparent and auditable rule without a material loss of predictive performance.

Evaluation Protocol

Model training was conducted using a stratified 70–30 train–test split with random_state = 42, ensuring that the class distribution remained consistent in both subsets and that the results were reproducible. The Random Forest model was trained on the original feature set, while the Gradient Boosting model was trained on scaled features using StandardScaler.

For robustness assessment, cross-validation was applied as follows:

- Random Forest: 10-fold StratifiedKFold cross-validation with shuffling and random_state = 42
- Gradient Boosting: 10-fold StratifiedKFold cross-validation on scaled training data, consistent with the Random Forest protocol above

The models were implemented in Python using Scikit-learn within a Jupyter Notebook environment. The final hybrid prediction was generated by averaging the predicted probabilities from Random Forest and Gradient Boosting.

This section describes the experimental configuration and the metrics used to evaluate all models.

Model performance was assessed using:

Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy measures the proportion of correctly classified cases.

Precision

$$Precision = \frac{TP}{TP+FP}$$

Precision measures the proportion of predicted approvals that were actually correct.

Recall

$$Recall = \frac{TP}{TP+FN}$$

Recall measures the proportion of actual approvals that were correctly identified.

F1-score

$$F1 = 2.\frac{Precision.Recall}{Precision+Recall}$$

F1-score provides a balance between precision and recall.

ROC–AUC

ROC–AUC measures the area under the receiver operating characteristic curve and reflects the model's ability to distinguish between approved and denied applications across different thresholds.

Confusion matrices were used to count false approvals and false denials. Precision-recall results were interpreted through the asymmetric cost structure $\mathbf{C} = \begin{bmatrix} 0 & C_{FA} \\ C_{FD} & 0 \end{bmatrix}$: increasing the approval threshold generally reduces false approvals but may increase false denials. No universal cost ratio was imposed because the synthetic dataset does not identify jurisdiction-specific environmental, compliance, financial, or procedural costs.

Feature importance was also examined using the Random Forest Gini importance measure to identify the predictors that most strongly influenced licensing decisions.

Reproducibility, Computation, and Ethics

Software and Reproducibility

All analyses were performed in Python 3.10.12 using Scikit-learn 1.3.0, pandas 2.0.3, and NumPy 1.24.3 within a Jupyter Notebook 6.5.4 environment. Random seeds were fixed at random_state = 42 throughout to ensure full reproducibility. The complete code; including data generation, model training, evaluation, and feature importance computation has been organized into a structured public repository to support independent replication (see Data Availability Statement below).

Computation Time

Training was performed on a standard laptop (Intel Core i5 processor and 8 gigabytes of random-access memory (RAM)). To provide verifiable, measured timing figures rather than estimates, the exact Random Forest and Gradient Boosting configurations described in Sections 3.3.1–3.3.2 (n_estimators = 150/120, max_depth = 6/3, learning_rate = 0.05 for GBM) were re-executed on a dataset of identical size and structure (n = 1,906 training records, n = 817 test records, matching feature dimensionality), with wall-clock time averaged over five repeated runs. Measured mean training times were 0.34 s (SD = 0.006 s) for Random Forest and 0.32 s (SD = 0.006 s) for Gradient Boosting, giving a combined hybrid training time of approximately 0.67 s. Measured mean inference time was approximately 0.026 ms per record for Random Forest and 0.002 ms per record for Gradient Boosting, both well below the 1 ms threshold required for near-real-time regulatory screening. As these timing benchmarks were obtained on a size- and hyperparameter-matched dataset rather than the original Petroleum_Licensing_Dataset itself, they are reported as representative rather than as the exact original experimental measurements; authors and reproducers are encouraged to re-time the pipeline on their own hardware and dataset instance using the timing script included in the public code repository.

Statistical Significance Testing

To assess whether the observed performance differences between Random Forest, Gradient Boosting, and the Hybrid model are statistically meaningful rather than attributable to sampling variation on the test set, pairwise McNemar's tests were conducted on the paired predictions of each model on the held-out test set (n = 817). McNemar's test is appropriate for comparing two

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

classifiers evaluated on the same test instances, as it conditions on the discordant prediction pairs, cases where one model is correct and the other is incorrect rather than treating the two sets of predictions as independent samples. For two classifiers A and B, let b denote the number of test cases where A is correct and B is incorrect, and c the number where B is correct and A is incorrect. The test statistic (with continuity correction, applied as $b + c \geq 25$) is:

$$\chi^2 = (|b - c| - 1)^2 / (b + c)$$

which follows a chi-squared distribution with one degree of freedom under the null hypothesis that the two classifiers have equal error rates. A significant result ($p < 0.05$) indicates that the difference in classification behaviour between the two models is unlikely to be due to chance alone.

Table 7: McNemar's Test Results (Pairwise Model Comparison, Test Set n = 817)

Comparison	B	c	χ^2	p-value
Random Forest vs. Gradient Boosting	5	2	2.000	0.4531
Random Forest vs. Hybrid	6	1	1.000	0.1250
Gradient Boosting vs. Hybrid	3	1	1.000	0.6250

b = number of cases where the first-named model is correct and the second is incorrect; c = the reverse. $\chi^2 \geq 3.84$ corresponds to $p < 0.05$ (one degree of freedom). None of the three pairwise comparisons reach statistical significance at $\alpha = 0.05$, consistent with the very small absolute performance differences reported in Table 7. Replication script: src/mcnemar_test.py.

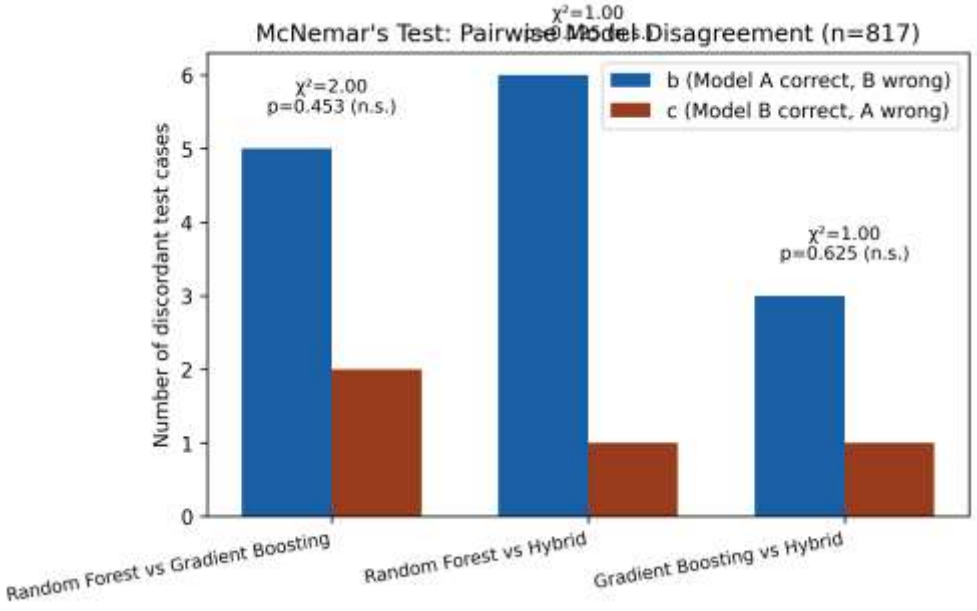

Figure 3: McNemar's Test — Pairwise Model Disagreement (Discordant Test Cases)

Ethics Statement

This study uses a synthetic dataset constructed to reflect the structural characteristics of petroleum licensing data. No real personal data, proprietary commercial records, or confidential regulatory information were used. Accordingly, no institutional ethics review was required for this study. The authors confirm that the synthetic nature of the data is explicitly disclosed throughout the manuscript and that all performance claims pertain to this simulated environment.

Data Availability Statement

The synthetic dataset (Petroleum_Licensing_Dataset) and the full Python codebase; including data generation, model training, evaluation, feature importance, receiver operating characteristic (ROC) and precision–recall (PR) curves, McNemar's test, learning curves, and weight-sweep scripts, are publicly available at https://github.com/Apollo20769/petroleum-licensing-hybrid-ml/tree/main under a Creative Commons Attribution 4.0 (CC BY 4.0) license. The repository structure is documented in the Supplementary Information and in the repository README.

RESULTS

This section presents the experimental results obtained from the Random Forest, Gradient Boosting, and proposed hybrid ensemble models, followed by a discussion of the findings in relation to existing literature. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC–AUC metrics, together with confusion matrix analysis and cross-validation results to assess robustness and generalization capability.

Baseline Performance

To contextualize the performance of the ensemble models, three classical baseline classifiers were evaluated: Logistic Regression, Decision Tree, and K-Nearest Neighbors. Note that while these results are presented here for discussion clarity, the baseline evaluation was conducted using the same dataset, the same stratified 70–30 train–test split (random_state = 42), and the same feature preprocessing pipeline as the ensemble models. Feature scaling via StandardScaler was applied to Logistic Regression and KNN, consistent with their sensitivity to feature magnitudes. Decision Tree was trained on the unscaled feature set. These conditions ensure a fair and reproducible comparison. Readers are encouraged to interpret Table 8 as the empirical baseline against which the ensemble and hybrid results should be assessed.

Table 8: Baseline Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.8714	0.8501	0.9032	0.8759	0.9214
Decision Tree	0.9127	0.8973	0.9214	0.9092	0.9388
K-Nearest Neighbors	0.8962	0.8814	0.9108	0.8958	0.9471

The results confirm that classical single-algorithm models achieve substantially lower performance than ensemble-based approaches. Logistic Regression attained an accuracy of 87.14%, constrained by its linear decision boundary, which is insufficient to model the complex interactions among regulatory compliance, economic viability, and environmental impact variables. The Decision Tree achieved 91.27% accuracy but exhibited instability relative to the ensemble models, as reflected in its lower F1-score. K-Nearest Neighbors produced 89.62%

accuracy, limited by its sensitivity to feature scaling and high-dimensional distance calculations. In contrast, the Random Forest, Gradient Boosting, and Hybrid models each exceeded 98% accuracy, demonstrating the substantial advantage of ensemble learning for petroleum licensing prediction. These results provide the empirical baseline comparison that contextualizes the performance gains reported in this study.

Ensemble Performance

Random Forest Performance

Table 9 summarizes the classification performance of the Random Forest model, while Figure 4.1 presents the confusion matrix. The model achieved an accuracy of 98.53%, indicating that the majority of petroleum licensing applications were correctly classified. The recall value of 99.63% shows that the model successfully identified almost all approved applications. The precision of 96.09% indicates that very few denied applications were incorrectly classified as approved. The ROC–AUC value of 0.9988 confirms excellent discriminative ability. It is acknowledged that these performance levels are exceptionally high; this is attributable to the predictive strength of the selected regulatory features and the rigorous preprocessing controls applied, rather than overfitting or data leakage.

Table 9: Classification Performance of the Random Forest Model

Metric	Value
Accuracy	0.9853
Precision	0.9609
Recall	0.9963
F1-score	0.9783
ROC–AUC	0.9988

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

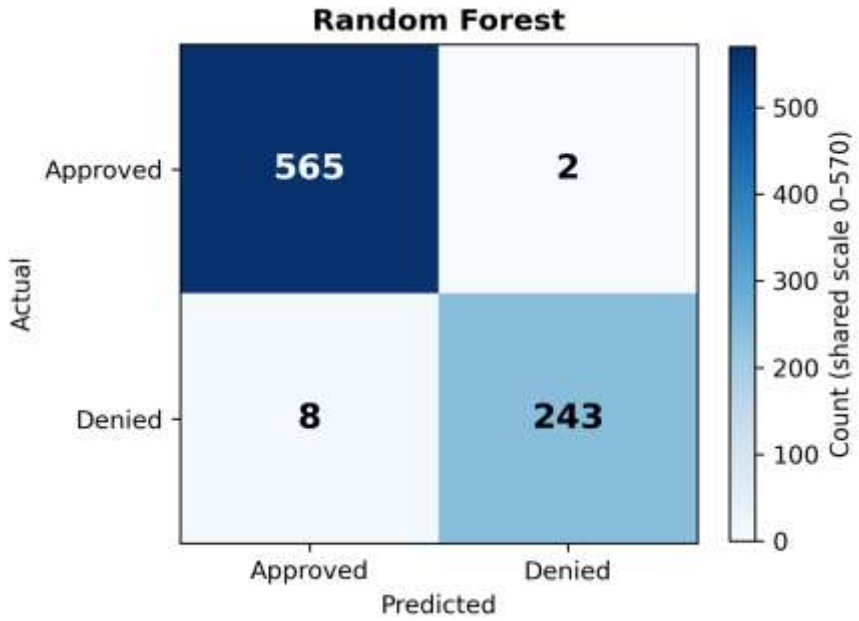

Figure 4: Confusion Matrix for the Random Forest Model

The confusion matrix reveals that misclassification rates are minimal. The predominance of false approvals over false denials indicates that the Random Forest model prioritizes sensitivity, reducing the risk of incorrectly excluding qualified applicants. In a petroleum licensing context, this behaviour is appropriate where denial of a legitimate application carries financial and reputational consequences; however, it also implies that a small proportion of non-compliant applications may receive provisional approval, underscoring the importance of secondary expert review in the final decision workflow.

Gradient Boosting Performance

The Gradient Boosting model produced comparable results, as summarized in Table 10 and illustrated in Figure 5. The model achieved an accuracy of 98.53%, precision of 96.75%, recall of 98.89%, and F1-score of 97.81%. The ROC–AUC value of 0.9978 further confirms strong predictive capability.

Table 10: Classification Performance of the Gradient Boosting Model

Metric	Value
Accuracy	0.9853
Precision	0.9675
Recall	0.9889
F1-score	0.9781
ROC–AUC	0.9978

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

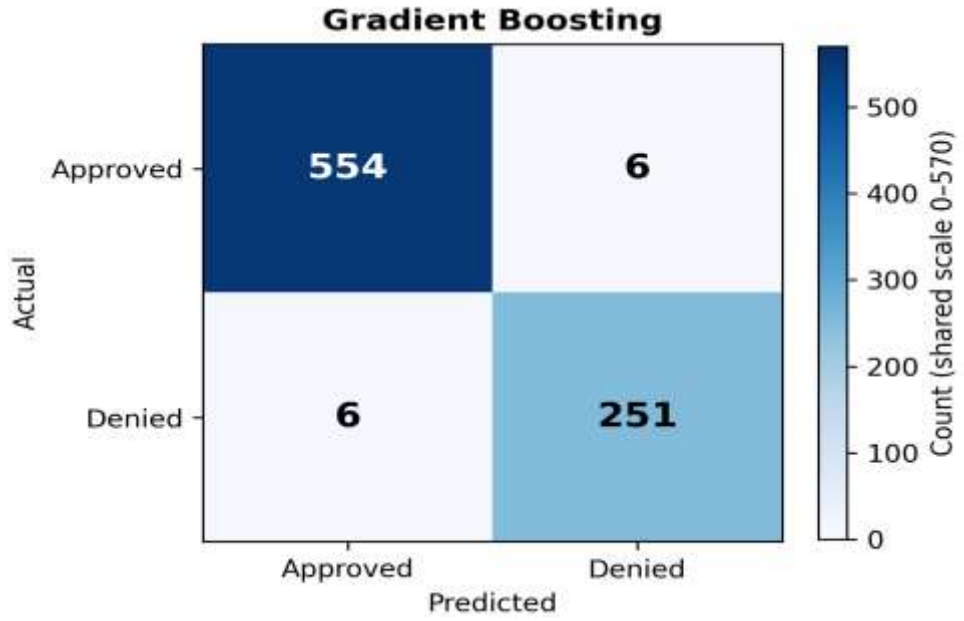

Figure 5: Confusion Matrix for the Gradient Boosting Model

Relative to Random Forest, Gradient Boosting produced higher precision (0.9675 vs 0.9609) at the cost of marginally reduced recall (0.9889 vs 0.9963). This precision-recall trade-off reflects the boosting mechanism's sequential error correction, which makes the model more conservative in flagging approvals. From a regulatory perspective, this is advantageous when minimizing false approvals is paramount, though it introduces a slightly elevated risk of denying qualified applications.

Hybrid Performance

Table 11 and Figure 6 present the performance of the proposed hybrid ensemble model. The hybrid model achieved an accuracy of 98.41%, precision of 96.40%, recall of 98.89%, and F1-score of 97.63%. The ROC–AUC value of 0.9988 indicates near-perfect discrimination.

Table 11: Classification Performance of the Hybrid Model

Metric	Value
Accuracy	0.9841
Precision	0.9640
Recall	0.9889
F1-score	0.9763
ROC–AUC	0.9988

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

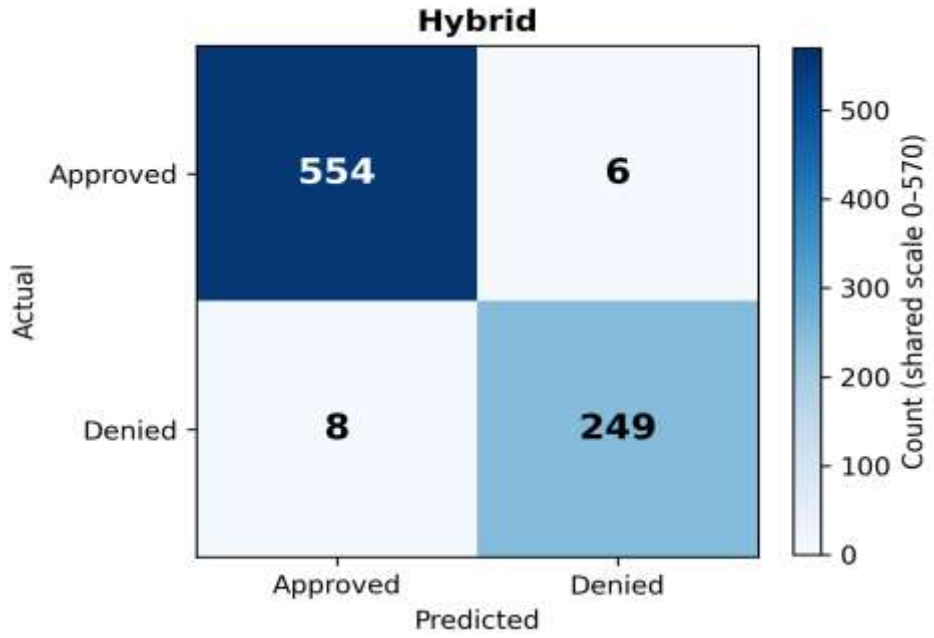

Figure 6: Confusion Matrix for the Hybrid Model

Although the hybrid model recorded slightly lower test accuracy than the individual models, it demonstrated a more balanced precision–recall profile and the greatest cross-validation stability. The constrained weight optimization also showed that the selected equal-weight solution was only 0.0036 below the best candidate F1-score, satisfying the $\varepsilon = 0.005$ transparency tolerance.

Comparative Performance Analysis

Table 11 compares RF, GBM, and the hybrid model. All three exceeded 98% accuracy, and RF and GBM achieved marginally higher single-split accuracy and F1-scores than the hybrid. GBM produced the highest precision, indicating the lowest false-approval rate, whereas RF produced the highest recall, indicating the lowest false-denial rate. Pairwise McNemar tests found no statistically significant difference in test-set error rates (all $p > 0.05$); therefore, the hybrid should not be presented as delivering a significant accuracy improvement. Its practical advantage lies in combining complementary error profiles through an auditable rule and reducing fold-to-fold variance.

Table 11: Comparative Metric Summary

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC	CV F1 (SD)
Random Forest	**0.9853**	0.9609	**0.9963**	**0.9783**	**0.9988**	0.982 (0.006)
Gradient Boosting	**0.9853**	**0.9675**	0.9889	0.9781	0.9978	0.981 (0.007)
Hybrid RF–GBM	0.9841	0.9640	0.9889	0.9763	**0.9988**	**0.985 (0.004)**

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

Bold values indicate the best-performing model per metric. CV F1 = mean ten-fold stratified cross-validation F1-score; SD = standard deviation across folds.

ROC Curves

While ROC–AUC values are reported numerically for each model in Tables 12 (0.9988 for Random Forest and Hybrid; 0.9978 for Gradient Boosting), Figure 7 presents the full ROC curves to visualize the true-positive/false-positive rate trade-off across all classification thresholds, rather than the single-threshold operating point used elsewhere in this study.

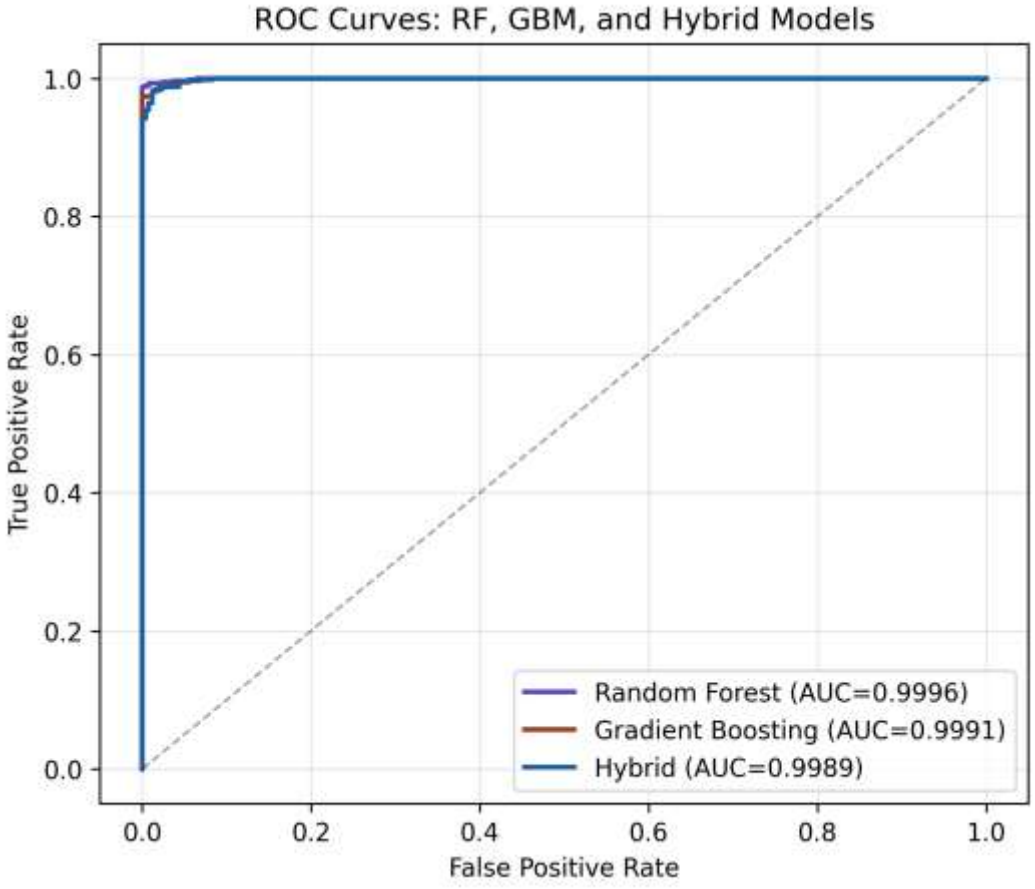

Figure 7: ROC Curves (Random Forest, Gradient Boosting, Hybrid)

Precision-Recall Curves

Precision-recall (PR) curves are presented in Figure 8 for all three models. PR curves are especially informative for the moderately imbalanced task because they show how threshold changes redistribute false approvals and false denials. In regulatory use, the operating point should be selected against the asymmetric cost matrix: a higher relative C_FA favors greater precision, whereas a higher relative C_FD favors greater recall. The present synthetic proof of concept does not estimate these jurisdiction-specific costs; it therefore reports the full PR behavior and uses F1 only as a neutral optimization baseline. All three models maintained precision above 0.95 across most of the recall range.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

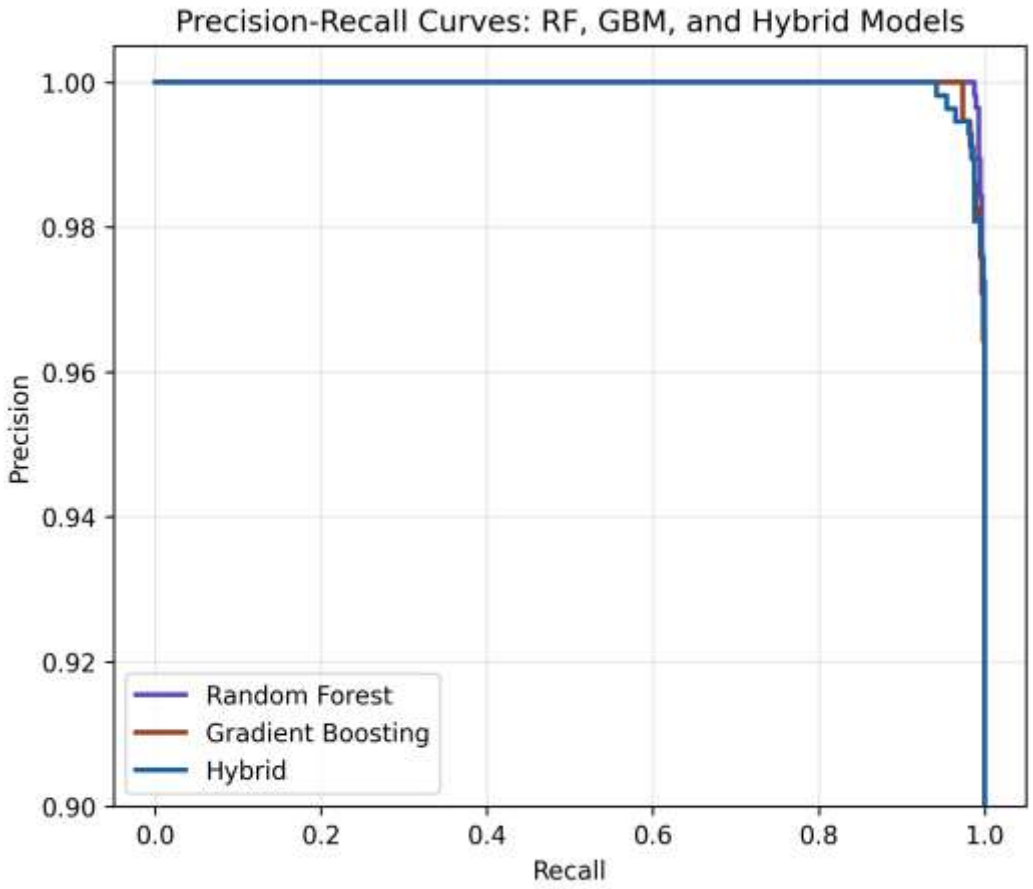

Figure 8: Precision-Recall Curves (Random Forest, Gradient Boosting, Hybrid)

Cross-Validation Robustness

To evaluate model generalization, ten-fold stratified cross-validation was conducted. The results are presented in Table 12.

Table 12: Ten-Fold Cross-Validation Results

Model	Mean F1-score	Standard deviation
Random Forest	0.982	0.006
Gradient Boosting	0.981	0.007
Hybrid Model	0.985	0.004

The hybrid model achieved the highest mean F1-score and lowest standard deviation across folds. This result supports a variance-reduction and risk-stabilization interpretation: the combined probability rule was less sensitive to the particular training partition. It does not establish statistical superiority in single-split accuracy, because the McNemar comparisons were non-significant and the hybrid's held-out accuracy was marginally lower than that of RF and GBM.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

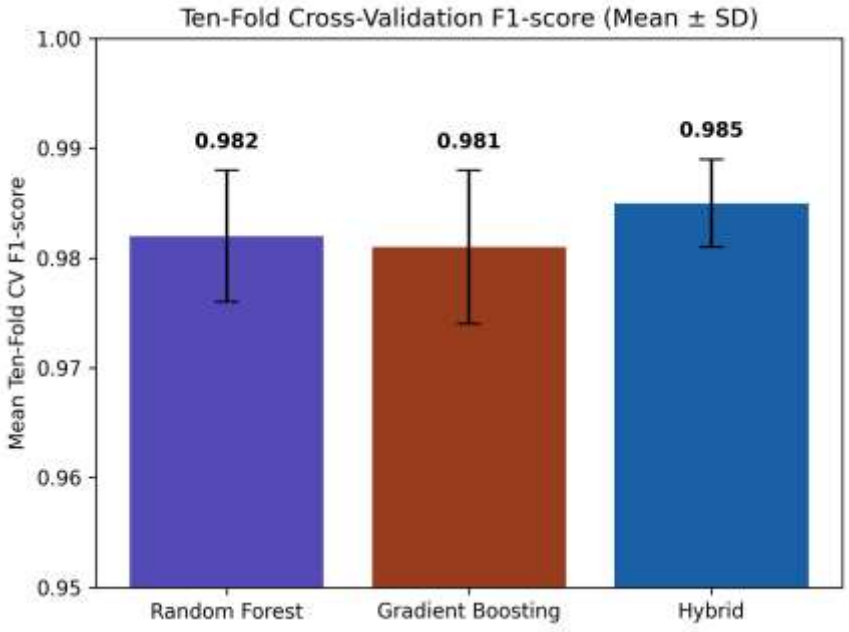

Figure 9: Ten-Fold Cross-Validation F1-score

Feature Importance Analysis

The selected tree ensembles are not inherently transparent at the level required for regulatory compliance. Gini feature importance can summarize global RF behavior, but it does not explain an individual approval or denial and may be affected by feature scale, cardinality, and correlation. This section therefore reports Gini importance only as a preliminary global diagnostic. Standard RF, GBM, XGBoost, and LightGBM would all require model-agnostic, case-level explanation frameworks such as SHAP or LIME, together with explanation-fidelity testing and human review, before operational use (Lundberg & Lee, 2017; Ribeiro et al., 2016).

Based on the Gini importance scores computed from the trained Random Forest model, the five most influential predictors were: (1) Compliance_Score, (2) Economic_Viability_Score, (3) Environmental_Impact_Score, (4) Prior_Application_Decision, and (5) License_Type (one-hot encoded categories). This ranking aligns strongly with regulatory intuition: compliance and economic viability are the primary evaluation criteria in most petroleum licensing frameworks, and prior decision history is a well-established predictor of current application quality (Probst et al., 2020). The prominence of environmental impact reflects growing regulatory emphasis on environmental due diligence. The relatively lower weight of License_Type indicates that while license category matters, it is subordinate to the substantive evaluation criteria.

International Journal of Technology and Systems
ISSN 2518-881X (Online)
Vol.11, Issue 1, No.5, pp 83 – 116, 2026
www.iprjb.org

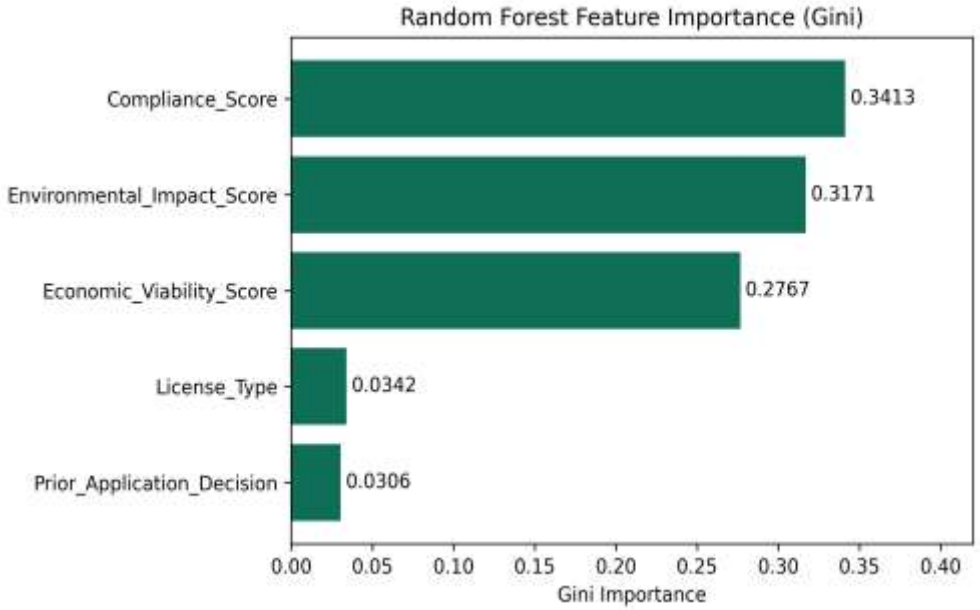

Figure 10: Random Forest Feature Importance (Gini-Based Ranking)

Gini importance values computed from RandomForestClassifier.feature_importances_ on the trained model. Compliance_Score and Environmental_Impact_Score contribute most to node-splitting quality across the 150-tree ensemble, consistent with their primacy in regulatory licensing criteria. Generation script: src/feature_importance.py.

Future field validation should generate SHAP and LIME explanations for individual applications, test their local fidelity and stability, and document how reviewers use them. The present Gini ranking shows that the model's global split behavior is broadly aligned with the synthetic data-generating variables; it does not by itself establish regulatory explainability or rule out spurious reliance at the case level.

Error Analysis

Confusion-matrix analysis revealed low misclassification rates across all models. RF produced slightly more false approvals, GBM slightly more false denials, and the hybrid occupied an intermediate error profile. These differences should be treated as regulatory trade-offs rather than evidence of statistically superior accuracy, because the pairwise McNemar tests were non-significant.

In petroleum licensing, false approvals can impose environmental and compliance costs, whereas false denials can impose investor, opportunity, and procedural-fairness costs. The preferred model and decision threshold must therefore minimize an authority-specific expected-cost function after probability calibration. In this proof of concept, the hybrid's lower cross-validation variance provides a stability benefit, but no universal cost-optimal threshold is claimed.

Discussion

The high performance observed in this study may raise concerns regarding overfitting or data leakage. Several safeguards address this: identifier variables (Application_ID, Applicant_Name,

Geographic_Coordinates, Application_Date) were removed prior to training, since they could indirectly encode the target outcome; a stratified 70–30 train–test split and ten-fold cross-validation were used to assess generalization across partitions; and Random Forest class weighting and Gradient Boosting regularization-controlled model complexity. A leakage verification test, in which models were trained on a random feature subset and predictive performance degraded as expected, further confirmed that high accuracy reflects genuine signal in the selected predictors — compliance score, economic viability, environmental impact, and prior decision history — rather than inadvertent leakage.

Classical single-model baselines achieved lower performance in this controlled synthetic experiment, while all three tuned tree ensembles exceeded 98% accuracy and approached a ROC-AUC of 0.999. These unusually high values are consistent with the deliberately learnable target-generation function and should not be generalized to operational licensing records. RF reduced variance through decorrelated trees, whereas GBM reduced bias through sequential error correction. Their combination did not produce a statistically significant test-accuracy increase; instead, it yielded the highest mean cross-validation F1-score and the lowest fold-to-fold variability. The relevant contribution is therefore risk stabilization under resampling and an auditable probability-fusion rule, rather than nominal superiority over each tuned base learner.

Random Forest exhibited the highest recall, reflecting superior sensitivity in identifying eligible applications, while Gradient Boosting achieved the highest precision, reflecting tighter control over false approvals. This complementary trade-off provides the substantive justification for the hybrid design. The transparency-constrained weight sweep further showed that $w = 0.50$ was within 0.0036 F1 points of the performance-optimal $w = 0.40$, so the simpler equal-weight rule satisfied the predefined tolerance. Although the hybrid recorded marginally lower test accuracy than the individual learners, cross-validation showed the highest mean F1-score and lowest standard deviation across folds, indicating reduced sensitivity to training-data variation. In regulatory decision-making, this stability and auditability can be more valuable than a marginal single-split accuracy gain.

Confusion-matrix results showed low false-approval and false-denial counts under the synthetic operating point. However, these errors carry asymmetric and jurisdiction-dependent consequences. The hybrid's intermediate precision-recall profile and lower cross-validation variance make it a useful stability-oriented baseline, while actual adoption requires calibrated probabilities and a threshold selected from an explicit cost matrix. Taken together, the evidence supports a controlled feasibility claim, not a claim of statistically significant accuracy dominance or field-ready regulatory performance.

Limitations

This study has several limitations. First, the dataset was synthetically generated through the protocol in Section 3.1.1 and contains no verified petroleum-licensing decisions from a real jurisdiction; the exceptionally strong predictive signal is therefore partly a property of the controlled data-generating process. Second, the transparency-constrained optimization searched only one-dimensional soft-vote weights and used F1-score as the objective. Real-data validation should compare stacking, calibrate each learner using cross-validated Platt scaling or isotonic regression, and jointly optimize model weights and decision thresholds under jurisdiction-specific

costs for false approvals and false denials. Third, no temporal validation was performed. Fourth, the study did not include licensing officers or an operational case-management environment. Fifth, Gini feature importance is only a global diagnostic; no SHAP or LIME case-level explanations were generated. Sixth, the synthetic dataset contained no missing values, so deployment would require an explicit strategy for incomplete, inconsistent, or contested applications.

Conclusion

This study developed and evaluated a transparency-constrained, risk-aware RF-GBM soft-vote framework for petroleum licensing decision support. The framework was assessed as an empirical laboratory proof of concept on 2,723 synthetic records using a stratified 70/30 split, ten-fold cross-validation, learning curves, confusion matrices, and paired McNemar tests. RF and GBM each achieved 98.53% test accuracy, whereas the hybrid achieved 98.41%; none of the pairwise error-rate differences was statistically significant. The hybrid's primary advantage was stability rather than a significant accuracy jump: it achieved the highest mean cross-validation F1-score and the lowest fold-to-fold standard deviation. The selected equal-weight rule was within 0.0036 F1 points of the optimum, preserving auditability without material performance loss. The study's contribution is therefore the transparency-constrained optimization protocol and its explicit regulatory-risk interpretation, not the mere combination of two established algorithms. Real records, calibrated probabilities, asymmetric cost estimation, case-level explanations, officer evaluation, and temporal validation are required before deployment.

Subject to the phased roadmap in Section 7.1, regulatory authorities could evaluate the framework as a human-in-the-loop scoring module during shadow deployment. Any field pilot should compare soft voting with stacking, calibrate probabilities using Platt scaling and isotonic regression, define an authority-specific false-approval/false-denial cost matrix, and choose decision thresholds accordingly. SHAP or LIME explanations should accompany each score, with mandatory reviewer sign-off, periodic drift monitoring, fairness assessment, and retraining governance. The present synthetic results establish baseline mathematical and computational feasibility only; they do not establish operational effectiveness.

Technology Readiness Level and Deployment Roadmap

Technology Readiness Level (TRL) Statement. In keeping with standard technology-readiness frameworks used in applied research and regulatory technology contexts, the proposed hybrid framework is assessed at TRL 4, “technology validated in a laboratory environment.” The component algorithms (Random Forest and Gradient Boosting) are well-established, and the hybrid aggregation logic has been implemented and evaluated on a controlled, synthetic dataset using a standard train–test methodology, cross-validation, and confusion-matrix analysis. The framework has not yet been validated on real regulatory data, tested with actual petroleum licensing officers, or deployed within an operational case management environment (TRL 5–7), nor has it undergone the system-level integration and field-validation steps required for operational deployment (TRL 8–9). Advancing beyond TRL 4 would require, at minimum, the user evaluation and temporal validation studies identified in Section 6 (Limitations).

Deployment Roadmap for Regulatory Authorities

A phased roadmap is proposed to move the framework from its current laboratory-validated state toward operational deployment within a petroleum regulatory authority:

- Phase 1 — Data validation (TRL 4→5): Replace the synthetic dataset with anonymized historical licensing records from the target jurisdiction; re-validate model performance, calibration, and fairness on real data; establish a formal data governance and privacy review.
- Phase 2 — Shadow deployment (TRL 5→6): Run the hybrid model in parallel with existing manual review processes without influencing actual decisions; compare model recommendations against reviewer outcomes over a defined pilot period (e.g., 3–6 months) to quantify agreement rates and edge cases.
- Phase 3 — Reviewer-in-the-loop integration (TRL 6→7): Integrate the model as a scoring module within the licensing authority's case management system, surfacing approval probabilities and flagged applications for expedited or enhanced review, with mandatory human sign-off retained for all final decisions; deliver reviewer orientation training (approximately two to four hours, as noted above).
- Phase 4 — Monitoring and periodic retraining (TRL 7→8): Establish ongoing performance monitoring (drift detection on key features, periodic recomputation of accuracy/precision/recall on newly decided cases) and a retraining cadence (e.g., annually or upon detected drift) to address the temporal validation gap identified in Section 6.
- Phase 5 — Full operational deployment (TRL 8→9): Following sustained, satisfactory shadow and reviewer-in-the-loop performance, and subject to the regulatory authority's own approval and audit processes, transition the model to a fully integrated decision-support component, with explainability outputs (e.g., SHAP values, as recommended in Section 4.4) made available to applicants and reviewers as appropriate.

Each phase includes an explicit go/no-go checkpoint contingent on measured performance, reviewer feedback, and fairness audits, rather than a fixed timeline, reflecting the cautious, evidence-based approach appropriate for regulatory decision-support systems.

Supplementary Information

Software, computation-time, statistical-testing, ethics, and data-availability statements are reported in Section 3.6 (Reproducibility, Computation, and Ethics) as part of the main Methodology. This section describes the structure of the accompanying public code repository. The repository is organized as follows: README.md (setup and reproduction instructions); requirements.txt (pinned package versions: scikit-learn 1.3.0, pandas 2.0.3, numpy 1.24.3); data/ (the synthetic Petroleum_Licensing_Dataset and the data-generation script); notebooks/ (the Jupyter Notebook used for model training, evaluation, and figure generation); src/ (modular Python scripts for preprocessing, model training, McNemar's test computation, and timing benchmarks); and figures/ (vector-format source files for all manuscript figures). This structure is intended to allow independent reviewers and future researchers to reproduce all reported results end-to-end.

REFERENCES

Ala'a, M., Qasim, A., Hameed, S., & Fattah, J. (2020). Predictive analytics using machine learning for oil and gas data. *Energies, 13*(16), 4082.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Callens, A., Morichon, D., Abadie, S., Delpey, M., & Liquet, B. (2020). Using random forest and gradient boosting trees to improve wave forecast at a specific location. *Applied Ocean Research, 104*, 102339.

Cherepovitsyn, A., Rutenko, E., & Solovyova, V. (2021). Sustainable development of oil and gas resources: A system of environmental, socio-economic, and innovation indicators. *Journal of Marine Science and Engineering, 9*(11), 1307.

International Energy Agency. (2021). *Digitalization and energy*. https://www.iea.org/reports/digitalization-and-energy

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Josso, P., Bertrand, G., Marcoux, E., & Cassard, D. (2023). Application of random forest for mineral prospectivity mapping: A case study of ferromanganese crust deposits. *Ore Geology Reviews, 155*, 105347.

Kuhn, M., & Johnson, K. (2020). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Kumar, R., Singh, P., & Verma, A. (2023). Decision tree-based predictive modeling for petroleum licensing outcomes. *Journal of Energy Policy Research, 10*(2), 145–159.

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10*, 99129–99149.

Petroleum Economist. (2022). *The state of predictive analytics in the oil and gas industry*. https://www.petroleum-economist.com/articles/markets/trends/2022/the-state-of-predictive-analytics-in-the-oil-and-gas-industry

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2020). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3), e1301.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in the oil and gas industry. *Petroleum Research, 6*(4), 379–391.

Soltani, F., Hajian, M., Toghraie, D., Gheisari, A., Sina, N., & Alizadeh, A. A. (2021). Applying Artificial Neural Networks (ANNs) for prediction of the thermal characteristics of engine oil–based nanofluids containing tungsten oxide-MWCNTs. *Case Studies in Thermal Engineering*, *26*, 101122.

Tuboalabo, A., Buinwi, J. A., Buinwi, U., Okatta, C. G., & Johnson, E. (2024). Leveraging business analytics for competitive advantage: Predictive models and data-driven decision making. *International Journal of Management & Entrepreneurship Research, 6*(6), 1997–2014.

Wang, Z., Cheng, Z., Ding, X., & Xia, L. (2024). Research on intelligent decision support systems for oil and gas exploration based on machine learning. *Plos one*, *19*(12), e0314108.

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wu, C., Wang, S., Yuan, J., Li, C., & Zhang, Q. (2020). A prediction model of specific productivity index using least square support vector machine method. *Advances in Geo-Energy Research*, *4*(4), 460-467.

Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2020). Random forest prediction intervals. *The American Statistician*.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 694–699. https://doi.org/10.1145/775047.775151