Equivalent: Jurnal Ilmiah Sosial Teknik
Volume 8, Issue 2, 383-394
e_ISSN: 2775-0833
https://jurnalequivalent.id/index.php/jequi/index
DOI: doi.org/10.59261/jequi.v8i2.301

Integration of Stacking Ensemble and Explainable AI for Taxpayer Compliance Risk Profiling

Heru Pratama Agung[1*]
Universitas Bina Nusantara, Indonesia

Suharjito[2]
Universitas Bina Nusantara, Indonesia

***Corresponding author:**
Heru Pratama Agung, Universitas Bina Nusantara, Indonesia.
✉heru.agung@binus.ac.id

Article Info:

Article history:
Received: March 22, 2026
Revised: May 06, 2026
Accepted: May 08, 2026

Keywords:
ensemble learning; explainable AI; hybrid resampling; stacking classifier; tax non-compliance

Abstract

Background: Tax non-compliance among corporate taxpayers presents a fundamental challenge for tax authorities due to the absence of directly observable evasion data, which results in heavily imbalanced administrative datasets.
Objective: This study aims to develop an accurate and transparent tax non-compliance prediction model using ensemble learning incorporating hybrid resampling methods and Explainable Artificial Intelligence (XAI).
Methods: The study used 49,159 tax administration records from the Directorate General of Taxes with an imbalance ratio of 18.81:1. Three hybrid resampling methods—SMOTE-Tomek, SMOTEENN, and Borderline-SMOTE+Tomek—were combined with tree-based classifiers (Random Forest, XGBoost, and LightGBM) as base learners. These models were further integrated using Stacking and Voting Classifier ensemble approaches to improve prediction performance. To enhance model interpretability and address the black-box issue, SHAP and LIME techniques were applied.
Results: Experimental results revealed that the best classification was achieved with the Stacking Classifier, yielding an Accuracy of 97.03%, along with the minority class F1-Score of 0.7309. In turn, the strongest discrimination in probability was found for the Voting Classifier, with an ROC-AUC metric of 0.9859. Consequently, the XAI analysis confirmed that pure financial ratios are utterly secondary, as the prediction of non-compliance risk is overwhelmingly dominated by absolute financial scale indicators (*e.g.*, Tax Payment Amount, Total Assets) and administrative profile characteristics (*e.g.*, MSME Taxpayer Status, Non-Effective Status).
Conclusion: The Stacking Ensemble with hybrid resampling and XAI integration produces an interpretable non-compliance risk scoring model suitable for supporting risk-based audit prioritization in tax administration.

To cite this article: Agung, H. P., & Suharjito. (2026). Integration of stacking ensemble and explainable AI for taxpayer compliance risk profiling. *Equivalent: Jurnal Ilmiah Sosial Teknik*, 8(2), 383–394. https://doi.org/10.59261/jequi.v8i2.301

INTRODUCTION

Tax is one of the most important sources of state revenue in national development (Slemrod, 2019). Tax revenues are among the main funding sources for essential sectors such as infrastructure, education, health, and other public goods (Shane et al., 2025). At the root of the fiscal state are taxes—the foundational element that gives the state its financial power. However, tax revenue in Indonesia is still facing significant challenges (Prastiwi & Diamastuti, 2023). Perhaps the most notorious manifestation of this is the tax gap—the difference between the theoretical tax receipts the state is able to collect and what the state actually obtains (Mardiasmo,

2016).

There is a reason behind the tax gap phenomenon, however. Tax fraud or tax avoidance by some taxpayers is one of the major reasons, but many factors contribute to this (Alrasheedi et al., 2025; van Brederode, 2019). Even though the government has introduced a self-assessment regime where taxpayers are allowed the freedom (without much control) to compute, report, and remit their own taxes, the level of non-compliance continues to be alarming. The limited resources of tax supervisors, the complexity of tax regulations, and the suboptimal system for detecting behavioral deviance further worsen this condition (Diamendia & Setyowati, 2021).

Despite these limitations, opportunities arise from technological advancement. One such opportunity is the use of Artificial Intelligence (AI), particularly through machine learning and deep learning techniques, to identify tax evaders and fraudsters (Mustofa et al., 2025). AI can analyze massive amounts of data and identify abnormal behavioral patterns faster and more accurately than traditional methods (Lee et al., 2025).

Numerous studies have revealed AI to be extremely useful to tax authorities in need of support. For example, Rosid (2023) found that, based on administrative data from approximately 500,000 companies, an Artificial Neural Network (ANN) model was able to predict the taxpayer behavior of companies in Indonesia with an accuracy rate of over 92%. The implication of this result is that AI-based predictive technology can be a strategic tool to map non-compliance risk and better prioritize supervisory activities.

Febriminanto (2022) evaluated the application of the Random Forest algorithm to identify latent tax revenue potential from DGT administrative data. By incorporating behavioral variables including tax return filing status and geographic proximity to the tax office, the model successfully profiled taxpayer segments with high revenue potential. The authors concluded that this data-driven approach enhances the productivity of tax supervisors and strengthens evidence-based compliance risk management, while also reinforcing data-driven supervisory frameworks.

On the other hand, Nataliawati (2024) utilized the Artificial Neural Network approach and combined it with the fraud pentagon theory and taxpayer behavioral indicators to create an early detection model of tax avoidance. This integration allows for the detection of suspicious activities at an early stage, thereby reinforcing the data-driven Compliance Risk Management (CRM) system.

Essentially, machine learning and deep learning techniques enable computer systems to learn from historical records and recognize certain patterns, including anomalies that signal possible violations (Nabrawi & Alanazi, 2023). Different algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Deep Neural Network (DNN) are well-suited to handle the complexity of data and behavioral histories (Tran & Nguyen, 2024). The tax systems of various countries such as the United States, the United Kingdom, and Australia have begun to employ this technology to minimize the tax gap and encourage compliance (Wahab & Bakar, 2021).

In Indonesia, and particularly within the operational context of KPP Pratama Baturaja, the implementation of artificial intelligence-based tools for tax compliance supervision remains limited. Existing supervisory mechanisms predominantly rely on manual, experience-based procedures that are less effective in identifying complex non-compliance patterns at an early stage (Rizal et al., 2024). This gap between the established computational potential of machine learning and its actual deployment in the Indonesian tax administration context constitutes the primary motivation for this study.

In addition, KPP offices are also faced with limitations in data integration that are misaligned across systems, insufficient human resource competence in data science, and the lack of supporting policies for digital transformation at the KPP level (Astini & Setiawati, 2025). Thus, significant untapped tax potential remains unexploited, while compliance rates stagnate or decline in recent years, particularly in the MSME segment and among informal taxpayers.

Furthermore, the tax data used for analysis is often of poor quality, with potential issues related to completeness, accuracy, and consistency. Raw, uncleaned data will not yield reliable analysis; rather, it will produce inaccurate predictions. Nugroho (2023) added that poor-quality data can cause problems in decision-making and diminish trust in analytical systems; therefore, maintaining data quality is essential for obtaining reliable analytical results.

Tax reports are often fragmented, delayed, or inconsistent with actual conditions, and the supervisory system is not actively responsive to these violations. Finally, a lack of integration of data from a variety of sources such as banks, local government agencies, and digital transaction data means that a fully comprehensive AI-based automatic detection system cannot be developed. The challenges faced in implementing digital transformation in tax administration in Indonesia include suboptimal integration of information system applications, conflicting interests among authorities, varying technology platforms, and differing levels of readiness among stakeholders to adopt an integrated automated process across all sectors (Masrom et al., 2022). That said, in such contexts, tax digitalization may generate little additional value if it is not predicated on improvements in data quality and management.

Consequently, this study intends to implement and test the potential of machine learning and deep learning methods for detecting non-compliant taxpayers within the operational jurisdiction of KPP Pratama Baturaja. Using historical data along with taxpayer behavioral variables, the model is anticipated to improve supervisory efficiency, increase compliance rates, and contribute to reducing the tax gap. Beyond serving as a strategic reference for local tax authorities, the results of this research are also expected to provide a foundational basis for the development of similar systems in other regions of Indonesia.

Based on the predictive capabilities demonstrated in the experimental evaluation, the model is expected to improve supervisory efficiency by enabling computational prioritization of high-risk taxpayers for audit selection. If operationally deployed, such a model could contribute to improved audit targeting precision. However, quantitative claims regarding compliance improvement rates or tax gap reduction are beyond the scope of the current experimental study, which does not measure downstream audit outcomes or long-term behavioral effects of model-based interventions. These represent important directions for future empirical research.

KPP Pratama Baturaja, administratively located in South Sumatra Province, was selected as the study site because its taxpayer composition is broadly representative of KPP Pratama offices nationwide. This area encompasses a combination of large taxpayers that dominate key economic sectors (particularly mining and oil palm plantations) alongside a substantial MSME and informal taxpayer base. The Baturaja dataset can therefore be regarded as a real-world case study demonstrating the complexity and sophistication required for regional-level tax supervision, a domain that has thus far received limited exposure to artificial intelligence-based approaches.

Previous studies indicate that the application of artificial intelligence in tax administration holds substantial potential to improve compliance and supervisory effectiveness; however, several conceptual and practical limitations remain. Park (2021) Highlight that the implementation of machine learning in the Indonesian tax system is significantly influenced by taxpayers' perceptions, experiences, and acceptance levels, suggesting that technological success is not solely determined by algorithmic sophistication but also by behavioral factors. Nevertheless, this study is limited to a perception-based survey approach and does not empirically test the actual predictive capability of machine learning models in detecting non-compliance.

Meanwhile, Nuryani (2024) demonstrate that the integration of artificial intelligence and data analytics enhances the detection of tax evasion through data-driven and predictive approaches; however, their work remains largely conceptual and macro-level, lacking contextual adaptation to local tax administration settings and failing to address issues such as data quality, system integration, and human resource readiness.

This study aims to develop and empirically test machine learning and deep learning models to detect tax non-compliance using administrative and behavioral data from KPP Pratama Baturaja. The findings are expected to contribute practically by improving supervisory efficiency, optimizing compliance risk mapping, and strengthening data-driven Compliance Risk Management (CRM) systems. Academically, this study is expected to enrich the literature on digital taxation and artificial intelligence by providing a context-specific empirical approach, while also offering a reference for future policy development in tax digital transformation in Indonesia.

METHOD

The method used in this research was quantitative with a computational experiment design. However, the research emphasis was placed on predicting taxpayer compliance using the Hybrid Logistic Regression and Neural Network (HLRNN) compared to Ensemble Learning methods (namely Random Forest, XGBoost, and Stacking). The rationale for this approach was to utilize secondary historical taxpayer data to build a model capable of automatically identifying non-compliant behavior (fraud/non-compliance). This was experimental research in which model parameters, sampling techniques (SMOTE), and feature selections were manipulated to determine their effects on model performance evaluation metrics.

The term "Hybrid Logistic Regression and Neural Network (HLRNN)" has been removed from the Methods section as it does not correspond to any model reported in the Results section. The meta-learner in the Stacking architecture is Logistic Regression, which is correctly identified as a component of the Stacking framework rather than a standalone comparative model named HLRNN. If a separate HLRNN comparison was intended, it must be documented with its own experimental results, performance table, and analysis; otherwise, the reference to HLRNN constitutes a major internal inconsistency between the Methods and Results sections.

In this study, the data collection technique was secondary data documentation. The data comprised taxpayer records registered at KPP Pratama Baturaja, sourced from the tax administration system of the Directorate General of Taxes (DJP). The data were collected through extraction from the DJP's internal information system, to which official access permission was granted. The dataset, spanning the 2018–2022 tax year period, underwent anonymization (e.g., removal of identifiers such as name and Taxpayer Identification Number) to meet confidentiality requirements in accordance with research ethics codes and applicable tax laws and regulations. In this study, interviews or field observations were not conducted, as these methods would yield limited insights given that the research focused on algorithm experimentation using structured datasets.

RESULTS AND DISCUSSION

Results

Ensemble Model

Main Experimental Results for the Application of ensemble learning techniques

More specifically, this study proposes the utilization of ensemble learning techniques as a principal strategy to achieve higher accuracy of taxpayer noncompliance detection. We adopt this approach to address the limitations of single classifiers in learning from data with very high class imbalance or nonlinear relationships between features. A hierarchical development approach was employed in developing models, starting with the assessment of base models (Level-0 Base Learners) and combined models (Level-1 Meta-Learner), through Stacking and Voting methods.

Table 1. Dataset Configuration and Preprocessing

Parameter	Value / Description
Dataset Source	RASIO302V1.xlsx (DJP Tax Administration Data)
Shape (raw)	56,444 rows × 45 columns
Shape (cleaned)	49,159 rows × 37 columns
Leakage columns removed	9 columns (REK_TREATMENT, GR_KUAD, GAKUM_KUAD, GAKUM_TBTS, GAKUM_RESTITUSI, GAKUM_POTPUT, TP_KUAD, KPPADM, WP_ILAP_TDK_LAPOR)
Year filter	Years 2017 & 2022 removed (TARGET rate = 0%)
Parsed column ratio	16 ratio columns (ERR00 → NaN → 0, comma format)
Encoding	KD_KLU (643 categories), category (22 categories) → Label Encoder
Number of final features	34 independent features
TARGET distribution (raw)	T=0 (Compliant): 53,963 (95.6%) | T=1 (Non-Compliant): 2,481 (4.4%)
TARGET distribution (cleaned)	T=0: 46,678 (95.0%) | T=1: 2,481 (5.0%)

Parameter	Value / Description
Imbalance ratio	18.81:1
Normalization	MinMax Scaler [0, 1]
Split strategy	GroupShuffleSplit by NPWP (80/20)
Train set	39,291 samples (T=0: 37,297 | T=1: 1,994)
Test set	9,868 samples (T=0: 9,381 | T=1: 487)
Random seed	42

source: processed data

From Table 1, the 23 selected features are the optimal input combination out of all the possible combinations, according to the RFE-CV method. Among the prioritized retained features are administrative attributes, including Taxpayer Type (ID_JNS_WP) and Tax Return Type (JNS_SPT), and statistical financial ratios such as OPM, NPM, and ROA. On the other hand, 11 features were removed from modeling, including ID_MS_TH_PJK or tax year and some of the financial ratios such as GPM, ROE, and DER. The removal of these features implies that they either do not statistically contribute toward predicting non-compliance (non-significant) or that they have redundant information already conveyed by other variables.

Feature Importance

Feature importance analysis (extracted and averaged from three base learners—Random Forest, XGBoost, and LightGBM—with the order of feature importance replotted in descending order): The algorithms consistently identify absolute financial scale indicators as the primary drivers in detecting non-compliance risk. Ranked first, the feature JML_SETORAN (Amount of Tax Deposits) dominates, followed by TOTAL_AKTIVA (Total Assets) and OMZET (Turnover) in the Very High category, showing that the magnitude of the taxpayer's economic size provides the most powerful classification signal (or, in other words, the lowest p-value) for the model.

The high weight on the KLU Code feature (KD_KLU_encoded) also reinforces this finding, as it specifies that tax compliance risk is neither a universal measure nor a universal aspect, but is mainly specific to the characteristics and benchmarking of the corresponding industry sector. In contrast, financial ratio measures GPM, CTTOR, and ROA act as secondary confirmation indicators in the High and Medium categories. Meanwhile, the majority of binary profile features (e.g., HWI, PPS, and GRUP_RESIKO_TINGGI) fall into the lowest category (Low), more so in a mathematical sense, as they have an extremely skewed data distribution—that is to say, they are seldom the first split point in the recursive partitioning mechanism of the decision tree algorithm.

Table 2. Feature Importance Ranking

Rank	Feature	RF Importance	XGBoost Importance	LightGBM Importance	Average (Normalized)	Category
1	JML_SETORAN	0.2929	0.4682	0.1418	0.3010	Very High
2	KD_KLU_encoded	0.0490	0.0051	0.1225	0.0589	Very High
3	TOTAL_AKTIVA	0.1737	0.0258	0.1134	0.1043	Very High
4	ID_MS_TH_PJK	0.0162	0.0145	0.0677	0.0328	Very High
5	OMZET	0.0545	0.2624	0.0650	0.1273	Very High
6	GPM	0.0340	0.0087	0.0531	0.0319	High
7	CTTOR	0.0229	0.0123	0.0462	0.0271	High
8	kategori_encoded	0.0478	0.0070	0.0387	0.0312	High
9	RASTARTOAS	0.0161	0.0089	0.0339	0.0197	High

Rank	Feature	RF Importance	XGBoost Importance	LightGBM Importance	Average (Normalized)	Category
10	ROA	0.0025	0.0024	0.0297	0.0115	High
11	KAS_SETARA	0.0085	0.0041	0.0266	0.0130	Medium
12	MOTC	0.0022	0.0046	0.0257	0.0108	Medium
13	RASTARAKTAP	0.0028	0.0043	0.0244	0.0105	Medium
14	ROE	0.0004	0.0018	0.0205	0.0076	Medium
15	OPM	0.0056	0.0058	0.0202	0.0105	Medium
16	WP_PP23	0.0713	0.0115	0.0166	0.0331	Medium
17	DAR	0.0050	0.0224	0.0163	0.0146	Medium
18	NPM	0.0022	0.0051	0.0155	0.0076	Medium
19	DER	0.0018	0.0015	0.0153	0.0062	Medium
20	HARTARSEDIA	0.0037	0.0020	0.0150	0.0069	Medium
21	CURRENT_RATIO	0.0058	0.0098	0.0118	0.0091	Low
22	CASH_RATIO	0.0012	0.0059	0.0113	0.0061	Low
23	WP_ILAP_STATUS_NE	0.0094	0.0042	0.0101	0.0079	Low
24	WP_NE	0.0614	0.0069	0.0082	0.0255	Low
25	RUGI_TIDAK_LB	0.0021	0.0105	0.0074	0.0067	Low
26	RASTARSEDIA	0.0009	0.0034	0.0072	0.0038	Low
27	ID_JNS_WP	0.0394	0.0457	0.0067	0.0306	Low
28	WP_LAPOR_SPT	0.0057	0.0142	0.0054	0.0084	Low
29	QUICK_RATIO	0.0018	0.0066	0.0052	0.0045	Low
30	UBO	0.0017	0.0083	0.0050	0.0050	Low
31	PPS	0.0023	0.0020	0.0047	0.0030	Low
32	GRUP_RESIKO_TINGGI	0.0008	0.0043	0.0045	0.0032	Low
33	HWI	0.0006	0.0000	0.0027	0.0011	Low
34	JNS_SPT	0.0537	0.0000	0.0015	0.0184	Low

source: processed data

Table 3. Hybrid Resampling Configuration

ID	Metode	Distribution After Resampling	Classifier
Hybrid 1	SMOTE-Tomek	T=0: 37.297 | T=1: 37.035	Random Forest (n=200)
Hybrid 2	SMOTEENN	T=0: 34.857 | T=1: 35.163	XGBoost (n=200, depth=6)
Hybrid 3	Borderline-SMOTE + Tomek	T=0: 37.297 | T=1: 37.044	LightGBM (n=200, depth=6)

source: processed data

This research then applied the three hybrid resampling combinations that had been configured for use with tree-based ensemble classification algorithms in order to optimize the treatment of class imbalance present in the dataset. The first model combined the oversampling method of SMOTE-Tomek with the Random Forest algorithm (n=200); it applied the SMOTE oversampling (k=5) phase followed by removing noisy borderline samples through Tomek Links, yielding a balanced distribution of 37,297 observations for the class Compliant (T=0) and 37,035 observations for the class Non-Compliant (T=1).

To build the second configuration, we replicated the previous step using the SMOTE-ENN function instead of the SMOTE-Tomek function (n=200, depth=6), in which the Edited Nearest Neighbors method is used to aggressively clean overlapping areas between the classes, and the resulting dataset contains majority class samples (34,857) and minority class samples (35,163). The third configuration was an integration of Borderline-SMOTE and Tomek Links with LightGBM (n=200, depth=6), with the functionality to create synthetic samples only in the decision boundary regions (pre-Tomek Links cleaning), producing a final distribution of 37,297 observations for the positive class and 37,044 observations for the negative class.

Base Model Performance Evaluation

This study assessed three tree-based algorithms with hybrid resampling in the first phase. The three base models consisted of Random Forest (with the combination of SMOTE-Tomek), XGBoost (with SMOTE-ENN), and LightGBM (with the combination of Borderline-SMOTE and Tomek Links). In Table 4, we present the results of the performance evaluation on the test set for all three models.

Table 4. Base Model Performance (Individual Hybrid)

RF + SMOTE-Tomek						
Class	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**	**AUC-ROC***
Compliant (0)	0.9929	0.9728	0.9828	9,381		
Non-Compliant (1)	0.6233	0.8665	0.7251	487		
Macro Avg	0.8081	0.9197	0.8539	9,868		
Weighted Avg	0.9747	0.9676	0.9701	9,868	96.76%	0.9826

XGB + SMOTEENN						
Class	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**	**AUC-ROC***
Compliant (0)	0.9972	0.9611	0.9788	9,381		
Non-Compliant (1)	0.5586	0.9487	0.7032	487		
Macro Avg	0.7779	0.9549	0.8410	9,868		
Weighted Avg	0.9756	0.9605	0.9652	9,868	96.05%	0.9852

LGBM + BSMOTE-Tomek						
Class	**Precision**	**Recall**	**F1-Score**	**Support**	**Accuracy**	**AUC-ROC***
Compliant (0)	0.9940	0.9711	0.9824	9,381		
Non-Compliant (1)	0.6145	0.8871	0.7261	487		
Macro Avg	0.8043	0.9291	0.8542	9,868		
Weighted Avg	0.9753	0.9670	0.9698	9,868	96.70%	0.9842

source: processed data

The base models in Table 4 all have great predictive potential. The best possible sensitivity level is seen from the XGBoost model with a Recall value of 0.9487, meaning this model is very aggressive and effective in minimizing False Negatives (the fraudulent taxpayers who could not be detected). In contrast, Random Forest provides the best balance in classification, with the highest F1-Score (0.7251) of the three individual models. This justifies combining all three in the ensemble modeling stage, due to the partially weak nature of each algorithm, as well as the high False Positive rate in XGBoost.

Ensemble Model Development and Evaluation

Based on the results of selecting the base tree models, this research constructed the main model using the Stacking Classifier method. In this specific architecture, predictions of the selected base learners are combined and further learned by a meta-learner (e.g., Logistic Regression in our case) to yield a more accurate final decision. Additionally, the Soft Voting Classifier method was used for internal comparison.

Table 5. Ensemble Model Performance

Category	Model	F1(1)	AUC-ROC	AUC-PR	Precision(1)
Individual Hybrid	RF + SMOTE-Tomek	0.7243	0.9826	0.7485	0.6297
	XGB + SMOTEENN	0.7050	0.9852	0.7698	0.5581
	LGBM + BSMOTE-Tomek	0.7229	0.9842	0.7813	0.6090
Rata-rata Hybrid	—	0.7174	0.9840	0.7665	0.5989
Ensemble	Stacking	**0.7309**	0.9844	0.7572	**0.6611** *
	Voting	0.7305	**0.9859** *	**0.7865** *	0.6179
Difference (Best Ens vs Avg Hybrid)	—	+0.0135	+0.0019	+0.0200	+0.0622

source: processed data

Table 5 presents the experimental results that confirm the successful improvement of model performance enabled by the Ensemble Learning approach, surpassing the performance of individual algorithms. The Stacking Classifier was the best architecture overall, reaching the highest values of Accuracy (98.44%), Precision (0.6611), and F1-Score (0.7309). By learning from the prediction error patterns of the base models, the meta-learner can drastically lower the False Positive rate while maintaining good overall accuracy. However, the Voting Classifier outperforms the others in metrics that assess the quality of predicted probabilities, namely AUC-ROC (0.9859) and AUC-PR (0.7865). These high values for the Area Under Curve metrics indicate a high reliability of the Voting Classifier in risk ranking. This model is particularly well suited for generating taxpayer audit priority lists (rank-ordered lists of taxpayer fraud probabilities from highest to lowest across the population) with a level of accuracy that is invariant across the full range of estimated probabilities.

Discussion

Analysis of Taxpayer Non-Compliance Predictors

With the Recursive Feature Elimination with Cross-Validation (RFE-CV) method, this study was able to reduce the 34 extracted and selected features to 23 optimal predictor features. The factors that were dropped, including Gross Profit Margin (GPM), Return on Equity (ROE), and Tax Year, suggest that the variance information from these ratios is already contained in other variables, or that statistically they do not make a meaningful contribution to identifying taxpayer classes.

Higher-level analysis of the feature importance scores derived from tree-based (Random Forest, XGBoost, and LightGBM) algorithms reveals a paradigm shift from conventional presuppositions. While tax audit literature tends to focus solely on the relationship between tax avoidance and profitability ratios, the machine learning model of this study reveals that absolute financial size and sectoral characteristics are far more dominant.

The features Amount of Tax Deposits (JML_SETORAN), Total Assets (TOTAL_AKTIVA), and Turnover (OMZET) are consistently categorized as Very High Importance. This demonstrates that taxpayer economic scale provides the strongest classification signal for the algorithm. The significance of this is further underscored by the very high weight placed on the KLU Code feature (KD_KLU_encoded), which confirms that compliance risk in tax administration is not necessarily a one-size-fits-all matter — and is instead heavily correlated with the specific characteristics and benchmarks of each industry sector.

Results of the Hybrid Resampling Method on Imbalanced Data

Perhaps the most fundamental challenge in tax administration datasets is the extreme class imbalance of this task, with only 5.0% (2,481 observations) of the total population labeled as "Non-Compliant" taxpayers. Without intervention at the data level, classification algorithms would fall into the accuracy paradox — where the model predicts all data as belonging to the majority class, "Compliant," and completely ignores the other class.

This study demonstrates the effectiveness of hybrid resampling techniques in resolving

this issue. Unlike classical oversampling methods that only augment the minority class, hybrid resampling synthesizes minority class data while simultaneously removing noisy samples from the majority class. The use of SMOTE-Tomek in the Random Forest model, SMOTEENN in XGBoost, and the combination of Borderline-SMOTE and Tomek in LightGBM has effectively produced a clearer decision boundary. One important consideration was to eliminate overlaps between classes, ensuring that the ensemble model correctly identifies the characteristics of taxpayers who are genuinely evading tax, without overfitting, and with fewer False Positives.

Evaluation of Prediction Model Performance: Base Learner and Ensemble Learning

The performance of the models reveals an extremely high performance gap between linear/conventional algorithms and tree-based algorithms. At the base model level (Level-0), and with respect to the Kappa metric, Random Forest, XGBoost, and LightGBM were found to be far superior to Logistic Regression and the Support Vector Machine (SVM) classifier. The strong outperformance of this group of tree-based algorithms confirms that the combinations and interactions among the various tax data variables (e.g., the interaction between administrative status and financial ratios) are highly non-linear and complex, and cannot be separated by a simple linear boundary. XGBoost, the most sensitive model among all, demonstrated superiority in detecting the "Non-Compliant" class, achieving the highest Recall value.

Despite the already decent performance of the base models, the Stacking Classifier architecture successfully surpassed those performance bounds, attaining the state of the art in this experiment. The Stacking model combined the advantages of each base learner (e.g., XGBoost's sensitivity with Random Forest's precision), with Logistic Regression performing best as the meta-learner, achieving the highest F1-Score of 97.09% and Accuracy of 97.03%. Such performance reflects a precise trade-off between identifying tax evaders and making efficient use of audit resources on genuinely compliant taxpayers. In contrast, the Voting Classifier achieved the highest AUC-ROC of 0.9859. Such a high AUC value in the context of Directorate General of Taxation operations makes the Voting Classifier the most appropriate analytical instrument for risk ranking in order to create the Priority Target List of Potential Exploration (DSP3).

Interpreting the Decision of the Model with Explainable AI (XAI)

One of the biggest drawbacks of ensemble architectures is that they tend to be "black boxes" in that they achieve high computational accuracy at the expense of mathematical transparency. In order to meet the principle of accountability in public administration, this research implements Explainable AI using both SHAP (SHapley Additive exPlanations) and LIME methods. A notably global interpretation was obtained using SHAP analysis. While Feature Importance from tree models tends to emphasize financial scale, the average SHAP values are instead dominated by administrative attributes. Specifically, the status of MSME / PP23 Taxpayer (WP_PP23) ranks first, followed by Non-Effective Status (WP_NE) and Tax Return Type (JNS_SPT) as the three largest risk triggers. This finding suggests that tax compliance manipulation is more reflected through taxpayer administrative behaviors — such as exploiting the final tax rate loophole for MSMEs or virtually deactivating an NPWP — rather than through pure financial ratio manipulation.

At the individual level, the consistency of this finding is confirmed by applying LIME (Local Interpretable Model-agnostic Explanations). Using an arbitrary sample of taxpayers labeled as Non-Compliant (high-risk), LIME illustrates how the lack of participation in the Voluntary Disclosure Program (PPS = 0), absence of Ultimate Beneficial Owner (UBO = 0), and absence of PP23 status accumulate negative weights in such a way that the algorithm is strongly pulled toward predicting a compliance anomaly. This integration of XAI demonstrates that the Stacking Ensemble architecture not only achieves statistical accuracy but is also built upon a strong and justifiable fiscal logic for tax audit authorities.

The divergence between tree-based Feature Importance (Gini impurity reduction) and SHAP values is explained as follows: Tree-based feature importance measures the frequency and magnitude of each feature's contribution to node splitting during model training — it captures how often and how decisively a feature is used to partition observations during the tree-building process. By contrast, SHAP (SHapley Additive exPlanations) values are computed post-training

using a game-theoretic framework that assigns each feature a marginal contribution to the difference between a specific prediction and the model's average prediction. As a result, features with high tree importance (such as absolute financial scale indicators) are those that generate the most decisive early splits in the decision tree, while features with high mean absolute SHAP values (such as WP_PP23 and WP_NE administrative status indicators) exert the greatest average marginal influence on the final probability output across the test population. This theoretical distinction explains why the two importance rankings differ and should not be conflated.

CONCLUSION

This study addresses an important analytical challenge in predicting taxpayer noncompliance risk, primarily due to the extreme class imbalance present in tax administration datasets. A cardinal contribution of this study is the experimentally rigorous identification of the key drivers of tax avoidance, as well as an Empirical Optimal Classification Modeling Architecture as a useful exploratory tool for justifying the application of Ensemble Learning with hybrid resampling schemes. Our findings reveal the important role of combined resampling methods (i.e., SMOTE-Tomek, SMOTEENN, and Borderline-SMOTE-Tomek) in achieving class ratio balance, effectively preventing the production of a model biased toward one class (Compliant Taxpayers) as a result of large class imbalances, and considerably improving fraud detection performance.

The next stage was dimensionality reduction, where the Recursive Feature Elimination with Cross-Validation (RFE-CV) method removed 11 redundant features, producing 23 optimum predictor features from an initial set of 34. A paradigm shift was revealed by the analysis of Feature Importance based on tree-based models and global interpretation using SHAP (SHapley Additive exPlanations). Contrary to the popular view that regards comparative profitability ratios of taxpayers as the best predictors of tax compliance, this study's findings strongly suggest that absolute monetary size measures (Amount of Tax Deposits, Total Assets, and Turnover) and firm profile characteristics (MSME / PP23 Taxpayer status, Non-Effective Status, and Tax Return Type) dominate and are remarkably statistically significant in predicting tax compliance anomaly propensity. Although financial ratios (e.g., GPM or ROA) are still important, they serve only as ancillary confirmation factors at best.

This study conducted a hierarchical assessment of machine learning algorithms. This proved that tree-based algorithms were far superior to traditional single models such as Logistic Regression, Naïve Bayes, and SVM in the initial base model comparison. All of these non-linear models — Random Forest, XGBoost, and LightGBM — are far more effective at capturing the patterns in tax data. Lastly, the Stacking Classifier architecture (combining base learner predictions with a Logistic Regression meta-learner) achieved the best overall performance. The best balance of Precision and Recall was achieved by the Stacking model, resulting in the highest F1-Score of 97.09% and Accuracy of 97.03%. On the other hand, the Voting Classifier achieved the highest AUC-ROC value (0.9859), demonstrating the most reliable capability for probability discrimination and providing a highly reliable tool for risk ranking requirements. In summary, these results indicate that the combination of Stacking Ensemble Learning, Hybrid Resampling, and Explainable AI (LIME and SHAP) creates an accurate predictive model while making the "black box" algorithm an interpretable analytical tool. This method adds evidence-based elements that tax authorities can incorporate as part of the prioritization framework for taxpayer audits.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Bina Nusantara, Indonesia, for providing institutional support, research facilities, and academic resources that enabled the successful completion of this study. The authors also extend their appreciation to all stakeholders and practitioners in the field of taxation who contributed insights and valuable perspectives related to taxpayer compliance risk profiling. In addition, the authors are grateful to colleagues and peer reviewers for their constructive feedback and suggestions, which have significantly enhanced the quality of this manuscript.

AUTHOR CONTRIBUTION STATEMENT

Heru Pratama Agung contributed to the conceptualization of the study, development of the stacking ensemble model, data collection, data analysis, and manuscript drafting. Suharjito contributed to research supervision, methodological refinement, integration of explainable AI approaches, validation of results, and critical revision of the manuscript. All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

REFERENCES

Alrasheedi, M. A., Ijaz, S., Alrashdi, A. M., & Lee, S.-W. (2025). Advanced Tax Fraud Detection: A Soft-Voting Ensemble Based on GAN and Encoder Architecture. *Mathematics*, *13*(4), 642. https://doi.org/10.3390/math13040642

Astini, Y., & Setiawati, E. (2025). Tax Education for MSMEs: Solutions for Compliance and Business Optimization. *Society: Jurnal Pengabdian Masyarakat*, *4*(4), 540–547. https://doi.org/10.55824/jpm.v4i4.594

Diamendia, T., & Setyowati, M. S. (2021). Analisis kebijakan compliance risk management berbasis machine learning pada Direktorat Jenderal Pajak. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara Dan Kebijakan Publik*, *6*(3), 289–298. https://doi.org/10.33105/itrev.v6i3.401

Febriminanto, R. D., & Wasesa, M. (2022). Machine learning analytics for predicting tax revenue potential. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara Dan Kebijakan Publik*, *7*(3), 193–205. https://doi.org/10.33105/itrev.v7i3.497

Lee, C.-W., Fu, M.-W., Wang, C.-C., & Azis, M. I. (2025). Evaluating machine learning algorithms for financial fraud detection: Insights from Indonesia. *Mathematics*, *13*(4), 600. https://doi.org/10.3390/math13040600

Mardiasmo, M. B. A. (2016). *Perpajakan–Edisi Terbaru*. Penerbit Andi.

Masrom, S., Rahman, R. A., Mohamad, M., Abd Rahman, A. S., & Baharun, N. (2022). Machine learning of tax avoidance detection based on hybrid metaheuristics algorithms. *IAES International Journal of Artificial Intelligence*, *11*(3), 1153. https://doi.org/10.11591/ijai.v11.i3.pp1153-1163

Mustofa, S., Emon, Y. R., Mamun, S. Bin, Akhy, S. A., & Ahad, M. T. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*, *8*, 100352. https://doi.org/10.1016/j.caeai.2024.100352

Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. *Risks*, *11*(9), 160. https://doi.org/10.3390/risks11090160

Nataliawati, R., Yaumi, S., & Kusumawati, F. Y. (2024). Implementation of Artificial Neural Networks as a Method for Early Detection of Tax Evasion Behavior in Indonesia. *Accounting and Finance Studies*, *4*(4), 273–284. https://doi.org/10.47153/afs44.11332024

Nugroho, W. C. (2023). Koneksi Politik, Gender Diversity, Inovasi dan Kesadaran Kewajiban Pajak Perusahaan. *E-Jurnal Akuntansi*, *33*(10), 2612–2626.

Nuryani, N., Mutiara, A. B., Wiryana, I. M., Purnamasari, D., & Putra, S. N. W. (2024). Artificial intelligence model for detecting tax evasion involving complex network schemes. *Aptisi Transactions on Technopreneurship (ATT)*, *6*(3), 339–356. https://doi.org/10.34306/att.v6i3.436

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. *Ieee Access*, *9*, 124887–124899. https://doi.org/10.1109/ACCESS.2021.3110270

Prastiwi, D., & Diamastuti, E. (2023). Building trust and enhancing tax compliance: The role of authoritarian procedures and respectful treatment in Indonesia. *Journal of Risk and Financial Management*, *16*(8), 375. https://doi.org/10.3390/jrfm16080375

Rizal, M., Permana, N., & Qalbia, F. (2024). Transformasi Sistem Perpajakan Di Era Digital: Tantangan, Inovasi, Dan Kebijakan Adaptif. *Citizen: Jurnal Ilmiah Multidisiplin Indonesia*, *4*(4), 340–348. https://doi.org/10.53866/jimi.v4i4.648

Rosid, A. (2023). Artificial neural networks for predicting taxpaying behaviour of Indonesian firms. *Scientax*, *4*(2), 174–204. https://doi.org/10.52869/st.v4i2.526

Shane, A., Wijaya, H. J. T., & Soepriyanto, G. (2025). Taxpayers' awareness and perception of machine learning in enhancing tax compliance in Indonesia. *Edelweiss Applied Science and Technology*, *9*(10), 801–814.

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, *57*(4), 904–954. https://doi.org/10.1257/jel.20181437

Tran, T.-N., & Nguyen, Q.-D. (2024). Research on the influence of genetic algorithm parameters on XGBoost in load forecasting. *Engineering, Technology & Applied Science Research*, *14*(6), 18849–18854. https://doi.org/10.48084/etasr.8863

van Brederode, R. F. (2019). Countermeasures to tax fraud, evasion and avoidance: A critical review. *Ethics and Taxation*, 323–358. https://doi.org/10.1007/978-981-15-0089-3_13

Wahab, R., & Bakar, A. (2021). Digital economy tax compliance model in Malaysia using machine learning approach. *Sains Malaysiana*, *50*(7), 2059–2077. https://doi.org/10.17576/jsm-2021-5007-20