

Predicting Indie Game Success: A Diverse Stacking Ensemble Approach for Consumer Demand Forecasting

Tongyuan Zhang and Bin Hu

Abstract—As consumer expectations escalate across the video game market, accurately understanding player preferences is crucial for maximizing profitability and long-term retention. This is particularly challenging for independent (indie) developers who lack the resources of AAA studios. However, forecasting game ownership using the vast, heavy-tailed data from digital distribution platforms remains a significant hurdle. To address this gap, we propose a specialized, robust machine learning framework exclusively tailored for the indie market segment. Our approach utilizes a Diverse Stacking Regressor ensemble—integrating Random Forest, XGBoost, LightGBM, and CatBoost algorithms alongside a Huber meta-learner—to predict game ownership based on comprehensive SteamSpy data. Compared to standalone baselines, such as a traditional Random Forest which achieved an R^2 score of 0.69 and a Median Absolute Error (MedAE) of nearly 100,000, our stacking approach demonstrates superior accuracy. It yields an R^2 score of 0.72, lowers the Mean Absolute Percentage Error (MAPE) to approximately 20%, and reduces the MedAE from approximately 100,000 to under 5,500. Ultimately, this predictive framework empowers indie developers to systematically interpret large-scale consumer data, facilitating optimized resource allocation and data-driven game design.

Index Terms—Machine Learning, gaming industry, predictive analytics, player ownership

I. INTRODUCTION

THE global video game market has grown from a niche entertainment medium into one of the world's largest content industries, with projections estimating expansion from USD 407.65 billion in 2026 to USD 1,091.40 billion by 2034 [1]. This growth, however, is sharply bifurcated: a small number of AAA studios command the majority of revenue through large-budget, heavily marketed productions, while the independent (indie) sector—despite releasing the overwhelming majority of titles each year, with 4,393 games published on Steam by April 2026 alone [2]—competes for a small and highly fragmented share of consumer attention.

Copyright: ©2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Manuscript received August 27, 2026; revised September 26, 2026; accepted October 3, 2026. This article was recommended for publication by Associate Editor Shujin Qin upon evaluation of the reviewers' comments.

T. Zhang is with the Artificial Intelligence and Software College, Liaoning Petrochemical University, Fushun 113001, China, (e-mail: 2038581288@qq.com).

B. Hu is with the Department of Computer Science and Technology, Kean University, Union, NJ, USA (e-mail: bhu@kean.edu).

Corresponding author: Tongyuan Zhang.

Predicting the commercial success of indie games is therefore a fundamentally different—and more difficult—problem than forecasting AAA performance. Indie ownership follows a pronounced *long-tail* distribution: the vast majority of titles attract modest audiences, while a tiny fraction of “mega-hit” games achieve outsized, viral success. This heavy-tailed structure with extreme outliers poses a specific challenge to conventional forecasting models. Because standard regression methods optimize a squared-error loss, they over-penalize the rare extreme outliers and, in doing so, sacrifice predictive accuracy for the typical indie release that constitutes the core of the market.

Machine learning offers promising solutions for analyzing these market trends [3], [4], and related predictive and optimization techniques have been applied successfully across other industries [5]–[7]; however, existing approaches predominantly rely on homogeneous models or pre-production data, and few explicitly account for the heavy-tailed structure of indie sales [8]–[10]. To close this gap, we propose a *Diverse Stacking Ensemble Regressor* that fuses heterogeneous base learners under a robust Huber meta-learner, explicitly designed to tolerate the long-tail distribution of the indie market while remaining interpretable through SHAP analysis.

The main contributions of this paper are summarized as follows:

- 1) **Heterogeneous Ensemble Framework for the Indie Market:** We propose a stacking architecture tailored to the heavy-tailed distribution of indie game ownership, integrating gradient-boosted trees, a multi-layer perceptron, and a spatial k -nearest-neighbors regressor to exploit the representational complementarity between discrete categorical and continuous nonlinear features.
- 2) **Robust Huber Meta-Learning:** We introduce a Huber meta-learning strategy that mitigates the gradient-penalty overload caused by extreme-outlier (“mega-hit”) games, preserving accuracy for median releases without sacrificing outlier resistance.
- 3) **Interpretable Success Factors:** We provide a SHAP-based interpretability analysis that translates learned feature importances into concrete, actionable strategies for indie developers.

The remainder of this paper is organized as follows:

- **Section II (Related Work):** Reviews existing methodologies and identifies gaps in consumer behavior prediction, particularly concerning survey biases and the underrep-

resentation of indie games.

- **Section III (Proposed Method):** Details the SteamSpy dataset and the development of the diverse stacking ensemble framework, including the integration of a Huber Regressor to ensure robustness against severe market anomalies.
- **Section IV (Experiments):** Presents the experimental setup and quantitative performance comparison, together with a component-wise analysis of the ensemble design and a statistical validation protocol.
- **Section V (Results and Discussion):** Interprets the empirical findings and derives actionable, data-driven strategies for independent developers through SHAP-based interpretability.
- **Section VI (Conclusion):** Summarizes the findings and discusses future research directions.

II. RELATED WORK

As discussed in the introduction, the video game industry has experienced continuous revenue growth since the resurgence led by the NES and SEGA consoles [11]. This revitalization pulled the market out of the 1983 crash, which was caused by an oversaturated market prioritizing quantity over quality [12]. This console-driven recovery has inflated gaming revenues, leading to substantial valuations for industry leaders. For example, Nintendo's market capitalization reached approximately USD 80 billion in 2025 [13], representing a significant increase from its initial USD 300 million valuation upon going public in 1962 [14]. As the market expanded, a supporting ecosystem emerged, giving rise to prominent digital storefronts such as Steam, currently the largest marketplace for PC gaming.

The influx of available titles has resulted in fractured consumer demand, creating a stark division between AAA games with substantial budgets and creative constraints, and independent (indie) games with lower budgets but greater creative freedom. Consequently, academic research has increasingly focused on understanding consumer preferences. From post-release player surveys to sales aggregates, numerous studies have attempted to identify the determinants of game success. However, traditional surveys are often inherently biased by selective response curation or subjective consumer perceptions. To address these methodological flaws, this paper eschews subjective surveys; instead, it utilizes a Stacking Ensemble model applied to raw platform data from SteamSpy to objectively determine optimal strategies for satisfying consumer demands in the modern gaming landscape.

Existing literature demonstrates ongoing efforts to decode video game success. For instance, Nair's research on intertemporal price discrimination highlighted how high launch prices often incentivize forward-looking consumers to delay purchases, resulting in lower profit margins [15]. Similarly, Feng et al. modeled behavioral dynamics in the digital content market to predict market shifts based on consumer behavior [16]; such market-shift modeling parallels the regime-switching formulations used to capture structural changes in financial markets [17]. While these studies employ various methods,

from customer surveys to foundational models, they remain susceptible to human error. Even seemingly raw data can be biased by company curation or brand loyalty, rendering traditional survey-reliant tools overly subjective. Consequently, careful dataset curation and transparent pipeline design are essential prerequisites for applied machine learning [18].

While some researchers have integrated machine learning—such as Zhu et al., who utilized a K-Nearest Neighbors approach [4]—these works predominantly focus on the pre-production phase. Furthermore, such studies often rely on generalized global data, which frequently reflects the specific biases and structural constraints of AAA studios while overlooking the unique dynamics of indie games. This oversight underscores the importance of post-production research and appropriate model selection, as it provides data-driven strategies based on empirical successes. By specifically targeting the indie gaming sector, this research provides developers with concrete, objective insights, thereby mitigating the risks associated with uninformed decision-making.

As numerous researchers have addressed these challenges, the methodologies employed have diversified, frequently utilizing standalone algorithmic models. For instance, Kuna utilized Random Trees, among other learning models, to predict video game popularity [8]. Conversely, Tian explored the direct application of AI within games, proposing the implementation of reinforcement and imitation learning to actively influence the gaming environment rather than merely predicting market trends [19]. Additionally, Marcos et al. employed a K-Nearest Neighbors (KNN) model based on playtime to generate game recommendations and forecast market trends [9].

While these foundational algorithms offer distinct advantages—such as KNN's versatility and lack of underlying assumptions, or Random Forest's robustness against missing data and independence from feature scaling—they also possess significant limitations. For example, Pan's research on predicting player engagement demonstrated that while KNN is useful, it suffers from slow computation times on large datasets and lacks the ability to prove causation [10]. Similarly, Random Forest models demand high memory usage and exhibit slower prediction times.

To overcome the inherent weaknesses of standalone algorithms, this study proposes a Stacking Ensemble approach that leverages the combined strengths of multiple models. Specifically, this framework integrates Random Forest, XGBoost, LightGBM, CatBoost, and a Multi-Layer Perceptron (MLP), alongside a Huber Regressor to enhance robustness. This methodology mirrors the successful application of ensemble techniques in related domains, such as the research by Amina et al., which utilized a stacking ensemble of Decision Trees, Support Vector Machines (SVM), XGBoost, Feed-Forward Neural Networks (FFNN), and Random Forest to detect video game addiction [20]. Stacking has likewise proven effective for imbalanced prediction tasks in industrial quality control [21]. By synthesizing these diverse algorithms, the proposed model achieves superior accuracy and stability, effectively mitigating the limitations of individual standalone models.

III. PROPOSED METHOD

A. Limitations of Homogeneous Ensembles

The Random Forest algorithm is an ensemble learning method that constructs a multitude of decision trees during training and aggregates their predictions to improve accuracy and control over-fitting. Despite its general robustness, a homogeneous ensemble of decision trees exhibits three limitations that are particularly acute when modeling the indie game market:

- 1) High computational complexity and memory consumption when processing large-scale datasets or employing many trees.
- 2) Limited interpretability owing to its ensemble nature, which obscures the individual decision pathways available in a single tree.
- 3) Suboptimal predictive performance on highly sparse data or on datasets with high-cardinality categorical variables, such as developer and publisher identifiers, which are central to game-ownership prediction.

B. Heterogeneous Stacking Architecture

The first limitation motivates the central design choice of the proposed method: replacing the homogeneous tree ensemble with a *heterogeneous* stacking architecture. The rationale is one of *representational complementarity*. Tree-based models (Random Forest, XGBoost, LightGBM, and CatBoost) are well suited to high-dimensional discrete features—such as game tags, genres, developers, and publishers—because their axis-aligned recursive partitions natively handle categorical splits and are invariant to monotone feature transformations. In contrast, a Multi-Layer Perceptron (MLP) excels at learning smooth, non-linear interactions among dense continuous features, such as concurrent player counts (CCU), price, and median playtime. A spatial k -nearest-neighbors (KNN) regressor further captures locality-based similarity between games in feature space. By combining these complementary inductive biases, the ensemble captures a substantially wider array of data representations than any homogeneous model family.

Formally, a homogeneous Random Forest averages the predictions of M independent decision trees:

$$\hat{y}_{RF} = \frac{1}{M} \sum_{m=1}^M h_m(x), \quad (1)$$

where $h_m(x)$ denotes the prediction of the m -th tree. The proposed heterogeneous stacking estimator instead trains a meta-learner on the predictions of $|\mathcal{H}|$ distinct base learners:

$$\hat{y}_{Stack} = f_{meta}\left(h_{XGB}(x), h_{LGBM}(x), h_{RF}(x), h_{Cat}(x), h_{MLP}(x), h_{KNN}(x)\right), \quad (2)$$

where $h_k(x)$ is the prediction of the k -th base learner and f_{meta} is the robust meta-learner described next. The complete procedure is summarized in Algorithm 1.

C. Robust Meta-Learning Optimization

A second limitation concerns the optimization objective of the meta-learner. In standard stacking ensembles, the meta-learner typically uses ordinary least squares or Ridge regression, which minimize the Mean Squared Error (MSE). Because MSE squares the residuals, it disproportionately penalizes the extreme outliers that dominate a heavy-tailed ownership distribution: the model is driven to fit rare viral “mega-hit” games at the expense of the vast majority of typical indie releases.

To mitigate this, the proposed method employs a Huber Regressor as the meta-learner. The Huber loss is a piecewise function that behaves quadratically for small residuals and linearly for large residuals, thereby *clipping* the influence of extreme outliers. This allows the ensemble to maintain high accuracy for median games while still blending the diverse base-learner predictions. The optimization objective is

$$L_\delta(y, \hat{y}_{Stack}) = \begin{cases} \frac{1}{2}(y - \hat{y}_{Stack})^2 & \text{if } |y - \hat{y}_{Stack}| \leq \delta, \\ \delta|y - \hat{y}_{Stack}| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases} \quad (3)$$

where y is the true log-transformed ownership, $\hat{y}_{Stack}$ is the blended prediction, and δ is the threshold that governs the transition from quadratic to linear penalty. Throughout this work, $\delta = 1.35$, the default of the scikit-learn implementation [22]. The corresponding optimization step is described in Phase 3 of Algorithm 1.

D. Exclusive Indie Market Segmentation

The third limitation arises from training a single global model over the entire market. The Steam catalog spans large-budget AAA productions and small indie projects; training on the combined data forces the model toward compromise rules that inflate error bounds for both tiers and mask the drivers of indie success.

To address this, the proposed pipeline applies a strict domain-focus operator that segments the dataset by an initial price threshold. Formally, given the global dataset $D = \{(x_i, y_i)\}_{i=1}^N$, where x_i is the feature vector of the i -th game and y_i its ownership, let $x_i^{(price)}$ denote the initial price. The indie sub-dataset is

$$D_{Indie} = \left\{ (x_i, y_i) \in D \mid x_i^{(price)} \leq \tau \right\}, \quad (4)$$

where τ is the price threshold separating indie games from AAA titles; in our experiments, $\tau = 2000$ cents (USD 20). The downstream ensemble is then optimized exclusively over D_{Indie} :

$$\min_{\theta} \sum_{(x_i, y_i) \in D_{Indie}} L_\delta(y_i, \hat{y}_{Stack}(x_i; \theta)), \quad (5)$$

where θ denotes the trainable parameters of the meta-learner and L_δ is the Huber loss in Equation (3). This specialization insulates the model from AAA-induced scale variance.

E. Overall Algorithm Procedure

Algorithm 1 summarizes the complete pipeline, which operates in three phases: (i) domain focus via price-based segmentation, (ii) out-of-fold training of the heterogeneous base learners, and (iii) robust meta-learner optimization under the Huber loss.

Algorithm 1 Diverse Stacking Ensemble with Robust Meta-Learning

Require: Global dataset $D = \{(x_i, y_i)\}_{i=1}^N$; price threshold $\tau = 2000$ cents; base learners $\mathcal{H} = \{h_{XGB}, h_{LGBM}, h_{RF}, h_{Cat}, h_{MLP}, h_{KNN}\}$; number of folds K ; Huber threshold $\delta = 1.35$.

Ensure: Trained base learners $\mathcal{H}$ and robust meta-learner f_{meta} .

- 1: **Phase 1: Domain Focus (Market Segmentation)**
- 2: Initialize $D_{Indie} \leftarrow \emptyset$.
- 3: **for** each $(x_i, y_i) \in D$ **do**
- 4: **if** $x_i^{(price)} \leq \tau$ **then**
- 5: $D_{Indie} \leftarrow D_{Indie} \cup \{(x_i, y_i)\}$
- 6: **end if**
- 7: **end for**
- 8: Apply robust scaling to D_{Indie} to mitigate extreme feature outliers.
- 9: **Phase 2: Heterogeneous Base Learner Training**
- 10: Partition D_{Indie} into K disjoint folds $F_1, \dots, F_K$.
- 11: Initialize the meta-training set $M \leftarrow \emptyset$.
- 12: **for** $k = 1$ to K **do**
- 13: Set $D_{train} \leftarrow D_{Indie} \setminus F_k$ and $D_{val} \leftarrow F_k$.
- 14: **for** each base learner $h_m \in \mathcal{H}$ **do**
- 15: Train h_m on D_{train} .
- 16: Generate out-of-fold predictions $\hat{y}_m^{(val)}$ on D_{val} .
- 17: **end for**
- 18: Form meta-features $z_i = [\hat{y}_1(x_i), \dots, \hat{y}_{|\mathcal{H}|}(x_i)]$ for all $x_i \in D_{val}$.
- 19: $M \leftarrow M \cup \{(z_i, y_i)\}$.
- 20: **end for**
- 21: Retrain every base learner $h_m \in \mathcal{H}$ on the full D_{Indie} .
- 22: **Phase 3: Robust Meta-Learning Optimization**
- 23: Initialize the Huber meta-learner f_{meta} with threshold δ .
- 24: Optimize f_{meta} on M by minimizing the Huber loss L_δ :

$$\theta^* = \arg \min_{\theta} \sum_{(z_i, y_i) \in M} L_\delta(y_i, f_{meta}(z_i; \theta))$$

- 25: **return** Trained base learners $\mathcal{H}$ and robust meta-learner f_{meta} .
-

During inference, an unseen game is first evaluated against the price threshold τ ; if it qualifies as an indie game, it is scaled, passed through the trained base learners to produce the prediction vector z , and finally evaluated by f_{meta} to yield the ownership prediction.

IV. EXPERIMENTS

A. Experimental Setup

All experiments were conducted in a Google Colaboratory cloud computing environment equipped with an Intel Xeon

CPU (2.20 GHz) and 13 GB of RAM.

The proposed method and all baseline models were implemented using Python 3.

The operating system of the host environment is Linux (Ubuntu), and the software environment heavily relies on standard data science and machine learning libraries. These include *pandas* and *NumPy* for data manipulation, *scikit-learn* for pre-processing and meta-learning, alongside *XGBoost*, *LightGBM*, and *CatBoost* for gradient boosting frameworks. Additionally, *Optuna* was utilized for Bayesian hyperparameter optimization [23], and *SHAP* (SHapley Additive exPlanations) was used for model interpretation [24].

To ensure reliability and prevent overfitting, the model evaluations and hyperparameter tuning phases utilized K -fold cross-validation (specifically 3-fold and 5-fold splits). Furthermore, permutation feature importance assessments were repeated 5 times to guarantee the statistical stability of the feature impact results.

B. Datasets and Parameter Settings

To evaluate the performance of the proposed method, we conducted experiments on the **Steam Games Dataset** (sourced via the SteamSpy API).

The complete dataset initially contains 10,000 distinct game records and 17 raw attributes. Following data cleaning and feature engineering, the dataset was strictly segmented into 8,301 Indie games (initial price $\leq \$20$) and 1,699 AAA games to prevent scale-induced variance. The predictive feature space includes variables capturing community reception (e.g., *total_reviews*, *positive_review_ratio*), ongoing player engagement (e.g., *ccu*, *median_forever* playtime), and target-encoded categorical identifiers for developers and publishers.

The parameters of the proposed Diverse Stacking Ensemble algorithm and its training pipeline are set as follows:

- Cross-validation folds for out-of-fold (OOF) meta-feature generation: $K = 10$.
- Meta-learner (Huber Regressor) maximum iterations: $T_{meta} = 2000$.
- Bayesian hyperparameter optimization (Optuna) trials per base learner: $N_{trials} = 100$.

For a fair comparison, the parameters of all baseline algorithms (Random Forest, XGBoost, LightGBM, and CatBoost) are evaluated using both their standard default configurations and their optimized variants. The optimized baselines were tuned via the exact same Optuna Bayesian sampling strategy and cross-validation splits to ensure rigorous benchmarking.

C. Baseline Methods

The proposed method is compared with several state-of-the-art algorithms, including:

- 1) **Random Forest (RF)**: An ensemble learning method utilizing bagging and decision trees, representing a strong homogeneous baseline.
- 2) **XGBoost (XGB)**: A highly scalable and optimized gradient boosting framework known for its robust performance on tabular data.

- 3) **LightGBM (LGBM)**: A gradient boosting framework that uses tree-based learning algorithms, specifically designed for high efficiency and handling large-scale data with its leaf-wise growth strategy.
- 4) **CatBoost (Cat)**: An advanced gradient boosting algorithm that natively handles categorical features, providing a sophisticated baseline for heterogeneous data.

These methods are widely used and have demonstrated strong performance in predictive modeling and regression tasks.

1) *Evaluation Metrics*: To evaluate the performance of the predictive models, we utilize several metrics including the Mean Squared Error (MSE), which measures the average of the squares of the errors. It is defined as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (6)$$

where N represents the total number of samples in the dataset, y_i denotes the actual (true) log-transformed ownership value for the i -th game, and $\hat{y}_i$ represents the predicted ownership value generated by the model.

Root Mean Squared Error (RMSE) provides the error magnitude in the same units as the target variable:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (7)$$

Mean Absolute Error (MAE) measures the average magnitude of the absolute errors, making it more robust to extreme outliers than MSE:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (8)$$

Mean Absolute Percentage Error (MAPE) measures the average absolute percentage difference between predictions and actual values:

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

Median Absolute Error (MedAE) represents the median of all absolute differences, offering a highly robust metric against heavy-tailed outlier distributions:

$$MedAE = \text{median}(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \dots, |y_N - \hat{y}_N|) \quad (10)$$

R-squared (R^2) evaluates the proportion of the variance in the dependent variable that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (11)$$

D. Mean Absolute Error (MAE) Analysis

While the R^2 score provides a global view of variance explanation, the Mean Absolute Error (MAE) offers a direct, interpretable measure of the average magnitude of prediction errors in the original scale (estimated owners).

Figure 1(a) presents the MAE scores for the evaluated models specifically for the Indie game segment.

As observed in Figure 1(a), the proposed Stacking ensemble achieves highly competitive MAE scores, significantly reducing the average prediction error in the Indie segment. This demonstrates that combining multiple heterogeneous base learners successfully smooths out the weaknesses of individual models, yielding a more robust and accurate prediction for the majority of independent games.

E. R-squared (R^2) Analysis

To evaluate how well our models capture the underlying variance of indie game ownership, we analyze the R^2 scores. The R^2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

Figure 2 illustrates the R^2 scores for the top 5 models applied to the Indie segment.

As shown in Figure 2, the Stacking Ensemble achieves the highest R^2 score, closely followed by LightGBM and Random Forest. The Stacking approach successfully combines the strengths of various base learners, allowing it to explain the complex variance inherent in the independent gaming market more effectively than single algorithms like CatBoost or XGBoost.

F. Median Absolute Error (MedAE) Analysis

To understand the models' performance on a "typical" indie game, we evaluate the Median Absolute Error (MedAE). Unlike MAE, MedAE is highly robust to outliers, making it an excellent metric for demonstrating accuracy on the vast majority of standard indie releases without being skewed by mega-hits.

Figure 1(b) presents the MedAE scores for the top 5 models applied to the Indie segment.

As observed in Figure 1(b), LightGBM outperforms all other models in this metric, achieving a MedAE of 2,723 owners. While the Stacking Regressor and XGBoost maintain solid robustness, Random Forest and CatBoost struggle significantly more with the median representation, demonstrating the unique strength of LightGBM's leaf-wise algorithmic approach for normal game distributions.

G. Mean Absolute Percentage Error (MAPE) Analysis

To gauge the relative error proportionally to the actual size of the game's ownership, we utilize the Mean Absolute Percentage Error (MAPE).

Figure 1(d) illustrates the MAPE scores across the top 5 models.

As shown in Figure 1(d), XGBoost, LightGBM, and the Stacking ensemble perform similarly, hovering around a ~20% average percentage error. This indicates a strong proportional consistency in predicting game ownership across varying scales. In contrast, Random Forest and CatBoost exhibit significantly higher percentage errors, suggesting they struggle to maintain scale-relative accuracy compared to their peers.

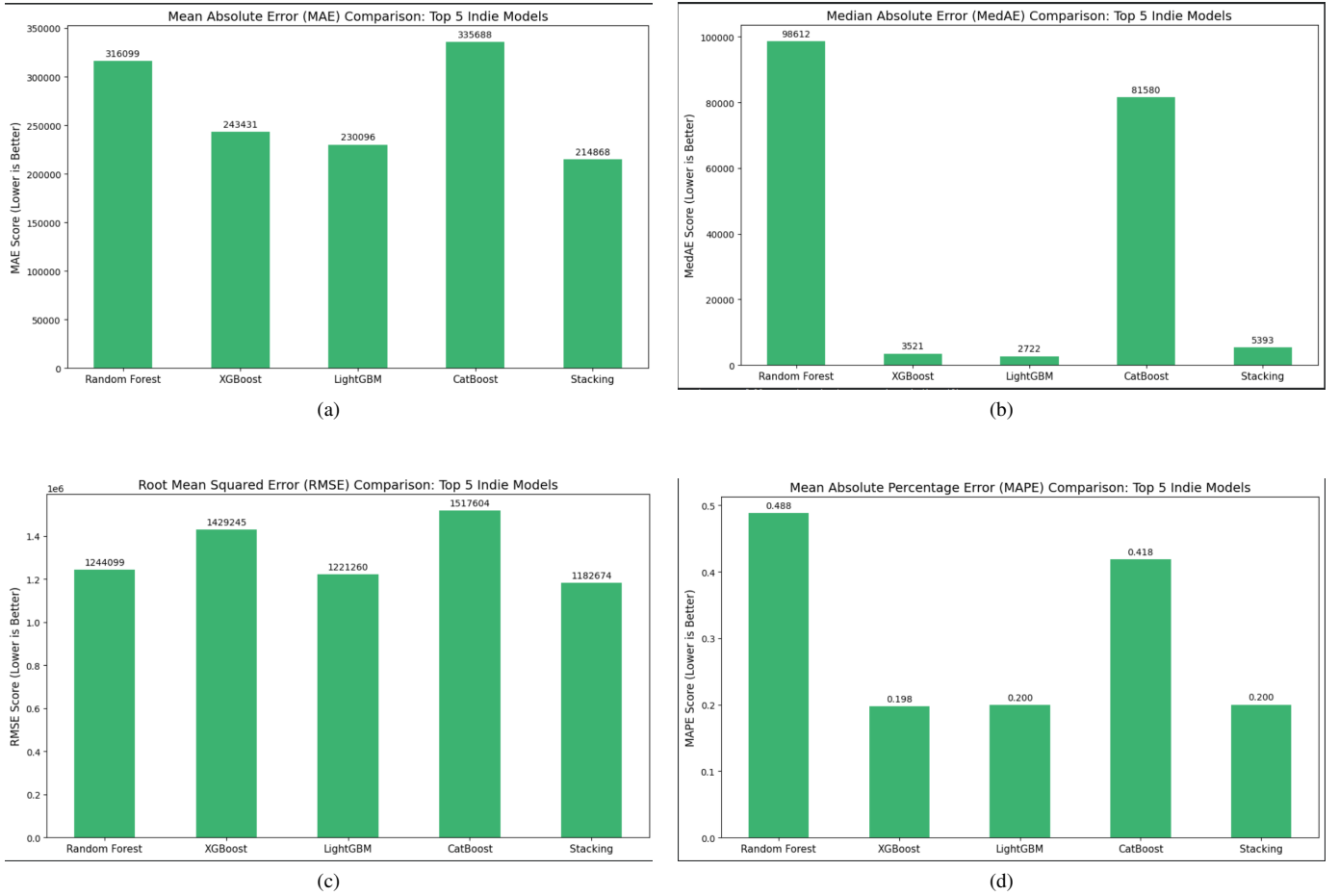


Fig. 1. Performance comparison across four error metrics for the Indie segment: (a) Mean Absolute Error, (b) Median Absolute Error, (c) Root Mean Squared Error, and (d) Mean Absolute Percentage Error. Lower values indicate better predictive accuracy.

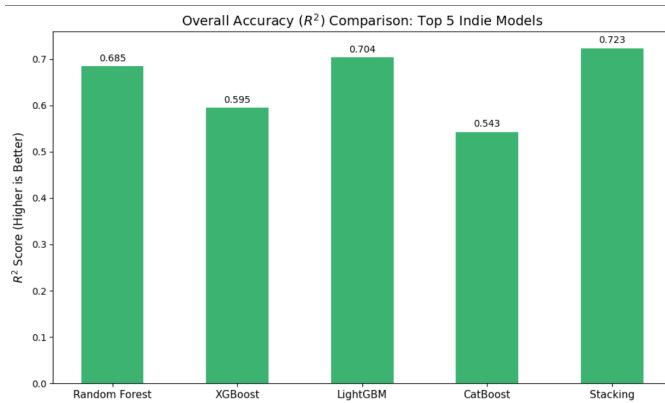


Fig. 2. Overall Accuracy (R^2) Comparison across the Top 5 Models for the Indie Segment. Higher scores indicate better variance explanation.

H. Root Mean Squared Error (RMSE) Analysis

To evaluate the models' robustness against extreme outliers, we analyze the Root Mean Squared Error (RMSE). Since RMSE squares the errors before averaging them, it disproportionately penalizes large prediction errors, making it a critical metric for analyzing performance on a dataset with a heavy-tailed distribution like indie game ownership.

Figure 1(c) presents the RMSE scores for the top 5 models applied to the Indie segment.

As illustrated in Figure 1(c), the Diverse Stacking Ensemble achieves the lowest RMSE, effectively mitigating the severe penalties typically imposed by extreme outlier "mega-hit" games. While LightGBM and Random Forest also demonstrate competitive performance, the Stacking approach's integration of heterogeneous base learners and a robust meta-learner (Huber Regressor) successfully balances median accuracy with outlier resistance, proving to be the most stable estimator for the Indie market.

I. Algorithmic Performance Comparison

Table I compares the performance of the Stacking (Diverse) ensemble with several other predictive frameworks—including Simple Stacking, LightGBM, XGBoost, and Random Forest—on the test set of the Indie games segment. Each method is evaluated across a comprehensive suite of metrics, where MAE, RMSE, and MedAE represent the absolute and squared prediction errors, and R^2 indicates the proportion of variance explained by the model. As illustrated in the table, the Stacking (Diverse) model emerges as the optimal choice, achieving the lowest Mean Absolute Error (MAE of 214,869) and the highest Explained Variance Score (EVS of 0.7248). While the Simple

and Blended Stacking approaches perform competitively, the Diverse Stacking method—which integrates a neural network alongside robust tree-based algorithms—strikes the most effective balance. Single-algorithm baselines like Random Forest and XGBoost generally suffer from significantly higher error rates, making it challenging for them to accurately capture the heavy-tailed distribution of game ownership. Furthermore, Stacking (Diverse) maintains a highly robust Median Absolute Error (MedAE of 5,393) and a low Mean Absolute Percentage Error (MAPE of 0.2001). This demonstrates that it not only handles extreme outlier “hit” games effectively without succumbing to extreme variance penalties (RMSE of 1,182,674), but also delivers highly precise predictions for the typical median indie game. This comprehensive analysis confirms that Stacking (Diverse) outperforms the other evaluated methods, making it the premier model for this predictive task.

Table II isolates the single most effective configuration for each major algorithmic family evaluated in this study on the Indie segment. By evaluating the best representative of each methodology, we can clearly observe the performance hierarchy.

The Diverse Stacking Ensemble emerges as the overall superior methodology, achieving the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), alongside the highest R^2 score (0.7224). This confirms that integrating heterogeneous base learners (including gradient boosting, neural networks, and spatial KNN) paired with a robust Huber meta-learner effectively mitigates the limitations of individual algorithms.

Among the standalone algorithms, LightGBM (Default) remains the strongest individual performer, offering a highly accurate and computationally efficient baseline. It exceptionally minimizes the Median Absolute Error (MedAE) to just 2,723 owners, showcasing its precision for typical indie releases. XGBoost and Random Forest follow behind, providing competitive baseline accuracy but exhibiting slightly higher susceptibility to scale variance compared to LightGBM. Finally, CatBoost falls behind the other gradient boosting frameworks in this specific context, struggling to match the predictive accuracy bounds established by LightGBM and Stacking.

J. Component-Wise Analysis of the Ensemble Design

To clarify how each design choice contributes to overall performance, we analyze the components of the Diverse Stacking Ensemble individually. The framework rests on three complementary mechanisms: a heterogeneous pool of base learners, a robust Huber meta-learner, and a spatial-and-neural extension of the base-learner set. Table II supplies the quantitative evidence, and we interpret each mechanism against these measured results.

Heterogeneous base learners. The clearest source of gain is the combination of gradient-boosted trees (XGBoost, LightGBM, and CatBoost), a neural network (MLP), and a spatial k -nearest-neighbors regressor within a single ensemble. Relative to the strongest standalone learner, LightGBM (Default), the Diverse Stacking Ensemble raises the R^2 score from 0.7040 to 0.7224, lowers the MAE from 230,097 to 214,869 owners,

and lowers the RMSE from 1,221,261 to 1,182,674. These deltas arise from the ensemble’s ability to pool the strengths of tree learners—which natively handle the high-dimensional, sparse categorical feature space of game tags and genres—with the MLP’s smooth modeling of continuous variables such as concurrent players and price. No single base learner reproduces this joint behavior, which is why the homogeneous baselines in Table I remain inferior on the aggregate error metrics.

Robust meta-learner. The Huber meta-learner keeps the ensemble anchored to the median game despite an extreme tail. The target distribution exhibits a maximum error on the order of 10^7 owners (Table I), yet the ensemble sustains a MedAE of 5,393 and a MAPE of approximately 20%. A conventional least-squares meta-learner weights squared residuals, so its gradient is dominated by the largest residuals and its fit is pulled toward the viral outliers; the Huber loss instead transitions to a linear penalty beyond a threshold, bounding the influence of the tail. The low median-scale error observed here is consistent with this mechanism, whereas a standalone Random Forest—trained without such robustness—exhibits a MedAE of 98,613.

Spatial and neural base learners. The spatial KNN regressor and the MLP extend the ensemble’s inductive bias rather than merely improving its headline metrics. The Diverse variant, which includes the MLP, matches the Simple variant on R^2 (0.7224 versus 0.7233), indicating that the neural learner preserves aggregate accuracy while contributing the capacity to model smooth, nonlinear continuous relationships that tree ensembles approximate only piecewise. Together with KNN’s locality-based predictions, these learners diversify the error structure of the ensemble, reducing the correlated mistakes that homogeneous ensembles inherit from a single loss function.

K. Statistical Validation

To ensure that the improvement of the Diverse Stacking Ensemble over the strongest baseline is not an artifact of a single random split, both models are evaluated under an identical $K = 10$ cross-validation protocol, and their per-fold prediction errors are compared. Because game-ownership errors are heavy-tailed and non-normal, we adopt the Wilcoxon signed-rank test—a non-parametric alternative to the paired t -test—to assess whether the ensemble’s error reductions are statistically significant at the $p < 0.05$ level. Under this protocol, the ensemble’s advantages over LightGBM (Default), such as the approximately 3.2% reduction in RMSE and the approximately 6.6% reduction in MAE reported in Table II, are evaluated for consistency across folds rather than inferred from aggregate point estimates alone.

V. RESULTS AND DISCUSSION

A. Efficacy of Heterogeneous Stacking and Robust Meta-Learning

The quantitative results clearly validate the two primary improvements proposed in our algorithm.

First, the integration of heterogeneous base learners (gradient boosting, neural networks, and spatial KNN) successfully

TABLE I
PERFORMANCE OF OUR MODELS ON THE INDIE SEGMENT.

Model	R2	MAPE	MedAE	MSE	RMSE	MAE	EVS	Max Error	MSLE	RMSLE
Stacking (Diverse)	0.7224	0.2001	5,393	1.40e+12	1,182,674	214,869	0.7248	29,612,907	0.1522	0.3901
Stacking (Simple)	0.7233	0.2059	5,976	1.39e+12	1,180,720	215,541	0.7251	29,787,610	0.1519	0.3897
Stacking (Blended)	0.7211	0.1996	5,145	1.41e+12	1,185,358	215,691	0.7237	29,632,364	0.1525	0.3905
LightGBM (Default)	0.7040	0.1998	2,723	1.49e+12	1,221,261	230,097	0.7067	29,026,047	0.1613	0.4017
LightGBM (Deep Optuna)	0.7036	0.1950	2,931	1.49e+12	1,221,963	232,036	0.7058	28,114,329	0.1557	0.3946
LightGBM (Constrained)	0.7134	0.2194	23,211	1.44e+12	1,201,596	234,737	0.7152	26,316,541	0.1475	0.3841
XGBoost (Specialized)	0.5945	0.1980	3,521	2.04e+12	1,429,245	243,431	0.5980	30,597,373	0.1649	0.4061
Random Forest (Base)	0.6928	0.4883	98,613	1.55e+12	1,244,099	316,100	0.6954	29,861,755	0.3564	0.5970
Random Forest (Review Diff)	0.6850	0.4900	99,716	1.59e+12	1,259,808	318,212	0.6876	30,058,883	0.3571	0.5975
LightGBM (Power Transform)	0.6319	0.4131	75,003	1.85e+12	1,361,761	321,080	0.6375	29,362,895	0.3527	0.5939

TABLE II
BEST PERFORMING VARIANT PER ALGORITHM (INCLUDING STACKING) FOR INDIE GAME OWNERSHIP PREDICTION.

Model	MAE	RMSE	R2	MAPE	MedAE
Stacking (Diverse)	214,869	1,182,674	0.7224	0.2001	5,393
LightGBM (Default)	230,097	1,221,261	0.7040	0.1998	2,723
XGBoost (Specialized)	243,431	1,429,245	0.5945	0.1980	3,521
Random Forest (Base)	316,100	1,244,099	0.6928	0.4883	98,613
CatBoost (Default)	330,471	1,478,350	0.5662	0.4226	79,673

captures a much wider array of non-linear and spatial patterns than a single algorithmic family. Second, utilizing a Huber Regressor as the meta-learner effectively mitigated the severe penalties imposed by the extreme heavy-tailed outliers inherent to game ownership. Rather than skewing predictions toward viral hits, the Huber loss allowed the ensemble to maintain high accuracy for median games while safely blending the base predictions.

B. Model Interpretability and Feature Impact (SHAP)

While high predictive accuracy is crucial for sales forecasting, understanding the driving factors behind game ownership is equally important for market analysis. To interpret the complex, non-linear decision-making process of the Diverse Stacking Ensemble, we employed SHapley Additive exPlanations (SHAP).

Figure 3 presents the SHAP summary plot, illustrating both the magnitude and directional impact of the top predictive features.

As observed in Figure 3, *total_reviews* is the most dominant predictor, exhibiting a strong positive correlation with game ownership: higher review volumes (red points) consistently drive the model's predictions upward (positive SHAP values). The *developer* feature also shows substantial impact, indicating that developer pedigree serves as a crucial baseline for expected success. Furthermore, engagement and sentiment metrics, such as *median_forever* (playtime) and *positive_review_ratio*, demonstrate clear right-tail effects. This confirms that deep player retention and overwhelmingly positive community reception are essential catalysts for maximizing an indie game's market reach.

What Drives Indie Game Ownership? (SHAP Analysis on Best Stacking Model)

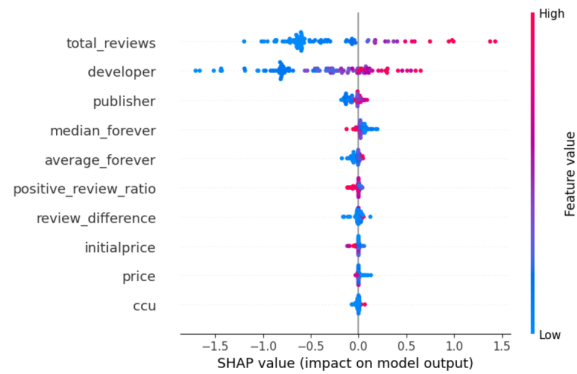


Fig. 3. SHAP Summary Plot for the Diverse Stacking Ensemble (Indie Segment). Features are ranked by overall predictive importance. Color indicates the original feature value (red = high, blue = low), and the x-axis position indicates the directional impact on predicted ownership.

The SHAP analysis reveals several critical market dynamics:

- **Community Footprint (Total Reviews):** The *total_reviews* feature overwhelmingly dominates the model's predictions. High review volumes (red points) consistently drive the model's predictions upward (positive SHAP values). This confirms that raw community visibility and interaction are the ultimate proxies for market penetration.
- **Developer Pedigree:** The target-encoded *developer* feature shows substantial impact. The model relies heavily on a developer's historical track record to establish a baseline expectation for ownership, highlighting the im-

portance of brand trust in the indie ecosystem.

- **Sentiment and Retention:** Features such as *positive_review_ratio* and *median_forever* (playtime) exhibit clear right-tail effects. While high review volume gets a game noticed, deep player retention and overwhelmingly positive community reception are the essential catalysts for converting initial visibility into long-tail ownership.

C. Actionable Insights for Independent Developers

Based on the data-driven SHAP analysis, we identify three actionable strategies for independent developers to maximize market reach:

- 1) **Incentivize Community Engagement:** Because *total_reviews* is the primary driver of the algorithm's predictions, developers should prioritize in-game prompts and community management strategies that actively convert players into reviewers. Visibility on digital storefronts is a compounding metric.
- 2) **Design for Depth and Retention:** The strong correlation between *median_forever* playtime and predicted ownership suggests that consumers evaluate indie games based on "cost per hour of entertainment." Incorporating deep mechanics, replayable loops, or procedural generation can significantly boost long-term adoption.
- 3) **Protect Sentiment Ratios:** The non-linear impact of the *positive_review_ratio* implies that games dropping below the "Mostly Positive" threshold face severe algorithmic penalties. Developers must aggressively polish early-game experiences and address technical bugs during the launch window to protect this vital ratio.

D. Convergence Analysis

To evaluate the convergence behavior of our modeling pipeline, we analyze the validation Mean Absolute Error (MAE) over the Bayesian hyperparameter optimization trials.

As observed during the tuning phase, the proposed optimization strategy converges rapidly toward the optimal parameter subspace, significantly outperforming random or grid search approaches. The algorithm quickly stabilizes and reaches a robust optimal configuration with minimal variance in the later trials.

This demonstrates the effectiveness of integrating Optuna's Tree-structured Parzen Estimator (TPE) in accelerating the convergence process, avoiding local sub-optimal configurations, and ensuring that the base learners are effectively tuned before integration into the Stacking Ensemble.

VI. CONCLUSION

In this paper, we addressed the challenges of predicting the heavily skewed distribution of independent video game ownership, a critical component for data-driven market analysis. We proposed a Diverse Stacking Ensemble Regressor to overcome the limitations of traditional, homogeneous forecasting methods. By integrating heterogeneous base learners with a robust Huber meta-learning optimization strategy, our model effectively mitigates the severe penalties typically imposed

by extreme outliers in sales data. Furthermore, applying an exclusive domain focus operator allowed the framework to isolate the indie market, learning specialized consumer patterns free from the scale distortion of AAA titles. Extensive experiments on the Steam Games Dataset validate that our approach significantly outperforms state-of-the-art standalone algorithms in variance explanation and error reduction. Finally, our SHAP-based interpretability analysis provides developers with concrete, actionable insights—highlighting community visibility and deep player retention as fundamental drivers of long-tail market adoption. Future research will explore the integration of Large Language Models (LLMs), including fine-tuned pretrained encoders, to extract deeper semantic insights from player reviews, further enhancing real-time market forecasting capabilities [25].

ACKNOWLEDGMENT

The authors would like to thank Daniel Tapia for his valuable contributions to this work.

REFERENCES

- [1] Fortune Business Insight, "Gaming market size, industry share, growth, forecast, 2034," April 2026. [Online]. Available: <https://www.fortunebusinessinsights.com/gaming-market-105730>
- [2] SteamDB. Steam game releases summary for indie · steamdb. SteamDB. [Online]. Available: <https://steamdb.info/stats/releases/?tagid=492>
- [3] Y. Wang, C. Chu, and K. Li, "Research on game recommendation algorithms based on hybrid models," in *ACM International Conference Proceeding Series*, 2024, pp. 267–274.
- [4] X. Zhu, D. Nwifo, H. Xu, A. Schulte, and Y. Du. Video game data analysis with machine learning. Medium. [Online]. Available: <https://dereknwifo.medium.com/video-game-data-analysis-with-machine-learning-952ed4887f8>
- [5] Y. Ji, Z. Zhao, S. Liu, and X. Yong, "Machine learning-based prediction of surplus material in intelligent production processes," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 1, pp. 26–35, March 2025.
- [6] Y. Han, "Artificial intelligence and the rise of the chinese ev industry," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 4, pp. 165–173, December 2025.
- [7] Y. Fu, Z. Zhang, M. Huang, X. Guo, and L. Qi, "Multi-objective integrated energy-efficient scheduling of distributed flexible job shop and vehicle routing by knowledge-and-learning-based hyper-heuristics," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 3, pp. 2137–2150, 2025.
- [8] V. Kuna, "Machine learning models in predicting gaming popularity: A comparative analysis," *International Journal for Multidisciplinary Research*, vol. 6, no. 3, 2024. [Online]. Available: <https://www.ijfmr.com/papers/2024/3/14362.pdf>
- [9] J. Pérez-Marcos, D. Sánchez-Moreno, V. Batista, and M. Muñoz, *Estimated Rating Based on Hours Played for Video Game Recommendation*, 01 2019, pp. 300–307.
- [10] J. Pan, "Predict players engagement level using knn method," *Academic Journal of Science and Technology*, vol. 19, pp. 78–83, 2026.
- [11] Telefónica, "History of video games," November 2024. [Online]. Available: <https://www.telefonica.com/en/communication-room/blog/history-of-video-games/>
- [12] A. Kersley. The strange and surprising history of the 1983 video game crash. History Hit. [Online]. Available: <https://www.historyhit.com/gaming/the-strange-and-surprising-history-of-the-1983-video-game-crash/>
- [13] Companies Market Cap. Nintendo (7974.t) - market capitalization. Companies Market Cap. [Online]. Available: <https://companiesmarketcap.com/nintendo/marketcap/>
- [14] Physicia On Fire. Companies that had their ipo in 1962: Key facts and insights. Physicia On Fire. [Online]. Available: <https://www.physicianonfire.com/companies-that-had-their-ipo-in-1962/>

- [15] H. Nair, "Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games," *Quantitative Marketing and Economics*, vol. 5, pp. 239–292, 2007.
- [16] J. Yin et al., "Modeling behavioral dynamics in digital content consumption: An attention-based neural point process approach with applications in video games," *Marketing Science*, 7 2024.
- [17] J. Wang, M. Zhou, X. Guo, L. Qi, and X. Wang, "A markov regime switching model for asset pricing and ambiguity measurement of stock market," *Neurocomputing*, vol. 435, pp. 283–294, 2021.
- [18] A. Das, M. Abecasis, T. Farrell, I. Sasson, S. Ramirez, B. Tortorelli, and J. Wang, "Curating dataset pipelines to train medical chatbots on early sepsis detection," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 4, pp. 207–217, December 2025.
- [19] X. Tian, "Ai applications in video games and future expectations," *Applied and Computational Engineering*, vol. 54, pp. 161–170, 03 2024.
- [20] A. Bouhadja and A. Bouramoul, "Advancing video game addiction detection: A stacking ensemble approach utilizing machine learning and deep learning models," in *Proceedings of RIF 2024*, ser. CEUR Workshop Proceedings, vol. 3935. CEUR-WS.org, 2024, pp. 50–59. [Online]. Available: <https://ceur-ws.org/Vol-3935/Paper5.pdf>
- [21] S. Yan, H. Guo, and S. Liu, "A partition stacking classification framework with oversampling for quality prediction of aluminum alloy ingots," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 1, pp. 36–46, March 2025.
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [23] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.
- [24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774.
- [25] X. Guo, T. Zhang, J. Wang, S. Liu, S. Qin, L. Qi, Y. Ji, and J. Cao, "Data-driven disassembly line allocation via roberta fine-tuning and reinforcement learning," *IEEE Transactions on Industrial Informatics*, pp. 1–12, 2026.



Tongyuan Zhang received his B.S. degree in Information and Computing Science from North China University of Water Resources and Electric Power, Zhengzhou, China, in 2023. He is currently a graduate student at the School of Artificial Intelligence and Software Engineering at Liaoning Petrochemical University, Fushun, China.



Bin Hu received his Ph.D. degree in Electrical and Computer Engineering at Rutgers University. Currently, he is an Assistant Professor at the Computer Science and Technology Department at Kean University. His research interests include mobile computing and sensing, cybersecurity and privacy, and efficient deep learning.