Vol. 11 No.2 Tahun 2025, pp 212-223
Jurnal Teknologi dan Manajemen Informatika
http:// http://jurnal.unmer.ac.id/index.php/jtmi
P-ISSN: 1693-6604 E-ISSN: 2580-8044

Market Value Tier Classification of Indonesian Football Players Using Ensemble Machine Learning and SHAP Analysis

Malfino Wildan Akhya[1], Cinantya Paramita[2], Pulung Nurtantio Andono[3]

Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Article Info

Article History
Received: 13-11-2025
Revised : 24-11-2025
Accepted: 17-12-2025

Keywords
Market Value Classification;
Ensemble Learning;
Indonesia Football Players;
SHAP

Corresponding Author
Cinantya Paramita,
Universitas Dian Nuswantoro,
cinantya.paramita@dsn.dinus.ac.id

ABSTRACT

The persistent discrepancy between actual transfer fees and the theoretical market values of football players highlights the need for a more objective and data-driven framework for player valuation. This study aims to classify the market value tiers of Indonesian Liga 1 players in the 2024/2025 season using an ensemble-based machine learning approach integrated with SHAP interpretability analysis. The dataset comprises 226 players with 27 attributes encompassing demographic, career, performance, physiological, and socio-economic dimensions. The research process involved secondary data collection, preprocessing, feature engineering, and percentile-based label construction, followed by model training using Random Forest, XGBoost, CatBoost, and a Stacking Ensemble. Experimental results show that the CatBoost model achieved the best performance, attaining an accuracy of 89%, a Macro-F1 score of 0.85, and an F1(High-Tier) of 0.78, demonstrating its robustness in handling heterogeneous and imbalanced data. SHAP analysis identified minutes played, age, and social media exposure as the most influential variables determining market value tiers. These findings demonstrate that combining ensemble learning with model interpretability can yield a transparent, adaptive, and practical framework for data-driven player valuation. The proposed approach provides actionable insights for football clubs and analysts in optimising player recruitment and developing fairer, evidence-based transfer strategies.

INTRODUCTION

The dynamics of modern football show that every club continuously pursues championship ambitions, with ownership and governance strategies playing a crucial role in shaping each club's risk profile and competitive orientation [1]. Furthermore, player quality remains a decisive factor in determining team success, as highly competent players contribute significantly to a club's overall performance and achievements [2]. Consequently, the transfer window has become a strategic period for clubs to strengthen their squads through player acquisitions, often involving substantial financial investments [3]. This aligns with findings that transfer spending is positively correlated with club performance and point accumulation in competitive leagues [4].

In the modern professional football ecosystem, the persistent mismatch between actual transfer prices and theoretical market values, has raised critical questions regarding market efficiency. Market value, commonly defined as an estimate of a player's fair value is often subject to significant distortion in practice [5]. Empirical evidence shows that the phenomenon of overpayment, where clubs purchase players far above their estimated market value, not only

undermines financial sustainability but also shows little correlation with consistent performance improvement [6]. Conversely, the presence of substantial "return premiums" for undervalued players indicates chronic underpricing when clubs release assets below fair value. This duality between overpayment and underpricing is not a temporary anomaly but a symptom of structural inefficiencies in conventional valuation mechanisms, which still rely heavily on intuition, public sentiment, and market speculation rather than data-driven assessment. Therefore, there is an urgent need for football clubs to adopt a more rational and objective valuation paradigm. In this regard, machine learning and econometric approaches that utilize granular datasets covering technical, physiological, and socio-economic indicators offer a transformative alternative [7]. Such models not only mitigate cognitive bias and information asymmetry but also enable dynamic and predictive value estimation, ultimately improving capital allocation and supporting clubs long-term competitive advantage.

Previous studies have explored various approaches to quantifying player market value. In the Indonesian context, Adiwiyana and Harymawan identified the determinants of professional football player valuation using multiple linear regression [8]. Their findings revealed that conventional performance metrics such as goals scored, minutes played, and starting appearances produced positive but statistically insignificant coefficients, suggesting that linear models may not sufficiently capture the valuation dynamics of a league characterized by heterogeneous player quality, uneven financial capacity, and inconsistent media exposure. Similar results were reported by Zulianto et al., who found that only goal scoring demonstrated statistical significance in Liga 1, while other performance indicators, despite being intuitively relevant, did not meaningfully influence player market value [9]. Together, these findings reinforce the argument that the Indonesian transfer market exhibits non-linear and interaction-driven patterns that traditional regression frameworks struggle to represent.

However, traditional regression-based approaches face structural limitations in modeling player valuation dynamics, especially in markets with skewed value distributions and complex interactions between performance, age, and contextual attributes. Machine learning techniques, particularly tree-based ensemble models such as Random Forest and Gradient Boosting, have demonstrated superior performance over linear regression by capturing complex non-linear relationships and higher-order interactions among predictors in football market valuation contexts [10], [11]. These ML-based models provide more adaptive frameworks that reflect real-world market asymmetries and have been shown to better support decision-making processes in transfer valuations and financial assessments within professional football analytics.

Beyond football analytics, classification-based machine learning models have been widely applied across various domains, demonstrating their methodological robustness and adaptability. For example, Michael and Rusman successfully implemented transfer learning with GridSearch optimization to classify defects in coffee beans, showing that classification models are effective for complex visual decision boundaries [12]. Similarly, Basri et al., developed a hybrid KNN–CNN model with Fourier Descriptor features to classify herbal leaves, further emphasizing the reliability of classification approaches for heterogeneous feature spaces [13]. In another domain, Nahak et al utilized convolutional neural networks to classify traditional Malaccan house types, demonstrating the capability of deep learning-based classifiers to differentiate between subtle architectural features [14]. A similar classification approach was demonstrated by Paramita et al., who employed the C4.5 decision tree algorithm to classify student characteristics, further reinforcing the versatility of machine-learning-based classification models across diverse application domains [15]. Collectively, these studies highlight that classification is an established and effective modelling strategy across diverse real-world applications, thereby supporting its use in this study to categorize football players into market value tiers rather than predicting precise numerical values.

Internationally, recent studies have increasingly employed advanced machine learning techniques. For instance, Behravan and Razavi developed an integrative model combining clustering algorithms with machine learning based on FIFA 20 and real-world statistics,

achieving significantly higher estimation accuracy than traditional models [16]. Consistent results were reported by McHale et al., who applied the Random Forest algorithm to European player data and observed substantial accuracy improvements [7]. Meanwhile, Sneha et al., a classification-based approach to profile elite player positions using anthropometric and motor fitness parameters, achieving 92% accuracy and an F1-score of 0.91, demonstrating the effectiveness of classification methods in identifying performance patterns [17]. However, most existing studies focus either on numerical market value estimation or on non-financial classifications such as player positions. A notable research gap remains: the direct application of classification models to categorize players into market value tiers as either an alternative or a complement to regression-based methods has been relatively underexplored, particularly in domestic leagues with distinct market characteristics like Indonesia Liga 1. This presents a methodological opportunity to better capture the inherently non-linear structure of player market value.

This study classifies player market value tiers in the 2024/2025 Indonesian Liga 1 using ensemble machine learning integrated with SHAP interpretability analysis. A binary-tier structure (high-tier vs. not-high-tier) was adopted to reflect Liga 1's concentrated value distribution and clubs' practical recruitment focus. Ensemble methods address player heterogeneity and data imbalance, while SHAP provides transparent feature contribution explanations. Evaluation uses accuracy, macro-F1, and F1 (high-tier) metrics to ensure comprehensive assessment, offering clubs data-driven insights for recruitment optimization and objective transfer decision-making.

METHODS

The research methodology serves as a systematic guideline in conducting the study, ensuring that each stage of the research is carried out in a focused manner and in line with the objectives to be achieved. The methodology is used to build a model for classifying the market value of football players based on ensemble learning. The research framework illustrating the stages of this process is presented in Figure 1. Broadly speaking, this research began with the collection of data on 226 players. The data was then divided into 80% for training and 20% for model testing. Next, preprocessing and feature engineering were carried out, followed by the formation of target labels based on market value percentiles to distinguish between High-tier and Not_High-tier classes. The final stage included ensemble model training, evaluation using performance metrics, and interpretability analysis using the SHAP method.

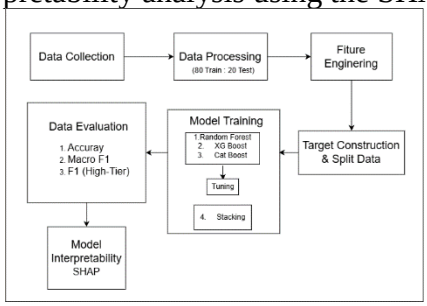

Figure 1. Research Flow Diagram

Data Collection

Secondary data was collected manually from four main sources: Transfermarkt, Soccer Wiki, Instagram, and Google News during the period 11–21 August 2025. The research focused on professional players participating in the 2024/2025 BRI Super League competition. The final dataset consisted of 226 player entries with 27 attributes presented in tabular format and grouped into five main categories: demographics, professional career, technical performance, physiology, and socio-economics. To maintain financial consistency, all player market values were converted from rupiah (Rp) to euros (€) using the daily average exchange rate during the data collection period. Table 1 shows the complete structure of the research dataset.

Table 1. Attributes used

Aspect	Attribute
Demographic	Age, Height, Weight, Nationality.
Professional Career	Current club, League, Years as a professional, Number of clubs, Contract remaining, Club value.
Technical Performance	Matches, Minutes played, Season goals, Season assists, National team goals, National team caps, National team awards, Club trophies, Penalties, Yellow cards, Red cards.
Physiological	Injuries (days).
Socio-Economic	Instagram followers, Google news exposure, Market value.

All numerical attributes are expressed in the International System of Units (SI): height (cm), weight (kg), time (minutes), age (years), injuries (days), and financial value (euro/€). The selection of attributes in this study was based on a synthesis of recent empirical studies that confirmed the relevance of each variable to determining the market value tiers of football players [5]. In terms of demographics and professional careers, variables such as age, height, nationality, club, league, and contract duration have a significant influence on player valuation. Sánchez et al. show that players aged 21–30 tend to have the highest market value, while the Parpa anthropometric study confirms that differences in height and physical characteristics across age groups correlate with performance capacity during the peak performance window [18], [19]. Sánchez et al., Bhilawa, and Fahriansyah emphasize that nationality influences market perception, with foreign players tending to receive higher valuations in domestic leagues [18], [20]. In the context of professional careers, Poli et al. identify experience and contract length as indicators of career stability related to transfer value, while Sánchez et al. and Satria highlight club reputation and league competition level as proxies for institutional prestige that directly imply player market value [21], [18], [22].

Furthermore, in terms of performance, physiological, and socio-economic aspects, a number of indicators contribute to the formation of a player's market value. Sánchez et al., Satria, and Utomo demonstrated that offensive performance, such as the number of goals and assists, has a positive correlation with market value, while Cai et al., showed that penalties and cards influence club's risk perception in the recruitment process [18], [22], [23]. Injury history has been proven to negatively impact performance stability and reduce player valuation [24]. On the other hand, digital popularity through the number of social media followers and the intensity of public coverage has a positive contribution to a player's commercial visibility and reputation, which ultimately increases market value [23], [21].

Data Preprocessing

The preprocessing stage is carried out to ensure that the data is of sufficient quality and consistency before being used in modeling. This process includes data cleaning, coding categorical variables, adjusting class weights, and dividing training and test data. Non-predictive attributes (Player_Name) are removed, while zero values in performance variables such as Goals_Season and Assists_Season are retained because they have valid statistical meaning. Categorical variables are encoded so that they can be processed by numeric-based models. In Random Forest and XGBoost, encoding is performed using one-hot encoding, while CatBoost processes categorical variables natively without explicit encoding. The Google_News attribute is encoded ordinally with a scale of Low = 1, Medium = 2, and High = 3.

The imbalance in class distribution (approximately 25% high-tier and 75% not_high-tier) is addressed through class weighting with the scale_pos_weight parameter in XGBoost and class_weights in CatBoost, which are adjusted based on the ratio of the number of samples in each class using the formula :

$$w_j = \frac{N}{K \times n_j}$$

where $w_j$is the weight for class j, N is the total number of samples, K is the number of classes (2), and n_j is the number of samples in class j .

The data was then stratified at an 80:20 ratio to maintain the class proportions in the training and testing subsets. The 80:20 data split using stratified train-test split follows current practices in football player valuation studies [18], [25], [26]. All of these preprocessing steps ensure that the dataset is clean, representative, and ready for use in the feature engineering and classification model training stages.

Fiture Enginering

The feature engineering stage was carried out to improve the models' ability to recognize non-linear patterns and complex interactions between variables through the formation of derivative attributes that are more representative of player performance. Four new features were constructed based on the rational domain of football, with the aim of expressing player productivity and experience in proportion to playing time and length of professional career. These derivative features were formulated as follows:

$$Goals\ Per\ Minute = \frac{Gol_{Musim}}{\max(Menit,1)}$$

$$Assists\ Per\ Minute = \frac{Assist_{Musim}}{\max(Menit,1)}$$

$$Contributon\ Per\ Minute = \frac{Gol_{Musim}+Assist_{Musim}}{\max(Menit,1)}$$

$$Long\ Career\ = 2025 - Tahun_{Pro}$$

These four features were selected because they conceptually reflect the actual performance and experience of players in standardized units of playing time, thus providing a fair basis for comparison between positions and clubs. The functional validity of these features was tested not only through the performance of ensemble models (Random Forest, XGBoost, and CatBoost), but also through interpretability analysis using SHAP to ensure the contribution of each feature to the classification results. This stage plays an important role in ensuring that the model is able to capture variations in player performance and experience in a more meaningful way without losing the semantic context of the original data.

Data Construction and Data Split

The target construction stage was implemented to transform the market value numerical variable into a categorical label representing the player's market value tier. This transformation aimed to adjust the research focus to the tier classification process (high-tier and not_high-tier), which was considered more relevant to player evaluation practices in the professional football industry. Categorization is performed using a percentile-based thresholding approach with a threshold at the 75th percentile. The use of the 75th percentile as a boundary is consistent with prior empirical studies that classify entities based on top-quartile thresholds, such as Kopka et al , who defined frontier firms using the upper 25% (75th percentile) of the distribution [27]. Players with market values at or above the 75th percentile are categorised as high-tier, while the rest are categorised as not_high-tier. This process is formulated as:

$$T_i = \begin{cases} 1, & MV_i \geq P_{75}(MV) \\ 0, & MV_i < P_{75}(MV) \end{cases} \quad (6)$$

where indicates the player category label for the i , MV_i is the market value in euros (€), and $P_{75}(MV)$ is the 75th percentile value of the Market Value distribution.

This process produces a binary target variable with an unbalanced distribution, namely approximately 25% High-tier and 75% not_high-tier. To reduce model bias towards the dominant class, class weighting is automatically applied based on the ratio of sample numbers

between classes. Next, the data is divided using the stratified train-test split method with a ratio of 80:20, in order to maintain a consistent class proportion in both data subsets. This process is implemented through the scikit-learn library with the parameter random_state = 42 to ensure consistent replication of the experiment.

Model Training

The model training phase in this study employed three ensemble algorithms random Forest, XGBoost, and CatBoost along with a Stacking Ensemble that combines XGBoost and CatBoost. Random Forest was used to reduce prediction variance through the bootstrap aggregation mechanism. XGBoost was selected for its ability to address class imbalance via the scale_pos_weight parameter and its strong built-in regularization scheme. Meanwhile, CatBoost was utilized due to its robustness on small datasets and its native support for handling categorical features through the ordered boosting strategy. Stacking was implemented to evaluate the consistency of predictions across algorithms rather than to enhance accuracy. This is because CatBoost demonstrated dominant performance, causing the meta-learner to naturally follow its prediction patterns.

Accordingly, stacking in this study serves primarily as a stability-validation mechanism rather than a performance-optimizing approach. To improve reproducibility, the training procedure incorporated hyperparameter tuning, particularly for XGBoost, using RandomizedSearchCV with Stratified 5-Fold Cross Validation and Macro-F1 as the optimization metric. Class imbalance was handled using scale_pos_weight, accompanied by threshold optimization within the range of 0.20–0.80 to maximize the high-tier F1-score. The configurations of CatBoost and Random Forest were determined through gradual parameter exploration. Table 2 shows a summary of the final hyperparameters is presented.

Table 2. Hyperparameters Used in Model Training

XGBoost (Tuned)		**CatBoost**		**Random Forest**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**	**Hyperparameter**	**Value**
n_estimators	300-900	iterations	300	n_estimators	100-400
learning_rate	0.03-0.1	depth	6		
max_depth	3-8	learning_rate	0.1		
subsample	0.7-1.0	loss_function	Logloss		
colsample_bytree	0.7-1.0	eval_metric	F1		
reg_lambda	0.0-5.0	class_weights	[1, 2.91]		
reg_alpha	0.0-1.0				
scale_pos_weight	2.91				

Given the relatively small sample size (226 players), mitigating overfitting was a critical component of the training process. In XGBoost, regularization using reg_lambda and reg_alpha, constraints on tree depth (max_depth), and sampling parameters (subsample, colsample_bytree) were applied to control model complexity. Stratified 5-Fold Cross-Validation was implemented to ensure stable performance across folds. CatBoost reduces the risk of target leakage through ordered boosting, while the use of class_weights in both XGBoost and CatBoost improves sensitivity to the minority class. In Random Forest, varying the number of trees (n_estimators) allowed the identification of a configuration that maintains stability without increasing variance excessively.

Hyperparameter tuning was essential because boosting-based models have a pronounced tendency to overfit on small datasets. Thus, tuning was performed not only to obtain the best predictive performance but also to ensure cross-fold stability. The combination of regularization, layered validation, and systematic parameter adjustment resulted in models that are more reproducible, stable, and well-adapted to the constraints of the dataset.

Model Evaluation

The evaluation stages were conducted to assess the model's performance in classifying football players into market value tiers that correspond to the empirical conditions of the data. Three main metrics were used in this study, namely accuracy, macro-f1, and f (high-tier). The Accuracy metric is used to assess the overall accuracy of the model in classifying all test data. However, because the class distribution in the dataset is unbalanced between the high-tier and not high-tier categories, this metric is supplemented by macro-F1, which calculates the harmonic mean between precision and sensitivity across all classes equally, so that the evaluation results are not biased towards the majority class. In addition, F1 (high-tier) is used specifically to assess the models ability to recognize players with high market values (minority class), which is considered most crucial in the context of player potential analysis and club recruitment policies.

The model evaluation was conducted using the hold-out validation method with an 80% training data and 20% testing data ratio in a stratified manner, so that the class distribution remained proportional. The metric formulation used is stated as follows:

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

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

$TP, TN, FP,$ and FN each representing the number of correct and incorrect predictions in the positive and negative classes. All Random Forest, XGBoost, CatBoost, and stacking ensemble models were evaluated using these three metrics to assess the stability and generalisation ability of the model against imbalanced class distributions in an objective and measurable manner.

Model Interpretability (SHAP)

Model interpretability analysis was performed using the SHapley Additive exPlanations (SHAP) method to explain the contribution of each feature to the classification results of player market value [28]. SHAP was chosen because it has a strong theoretical basis in cooperative game theory, which views each feature as a marginal contributor to the model output. This approach allows for a fair and transparent evaluation of feature influence. This method has been widely used in studies of complex model interpretability [29].

Mathematically, the SHAP value (ϕ_i) is expressed as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F is the feature set, S is the feature-free subset i, and $f(S)$ is the model prediction value for that subset. The value $\phi_i > 0$ indicates that the feature increases the probability of a player being in the high-tier category, while $\phi_i < 0$ decreasing it. SHAP analysis was conducted using two approaches. Global analysis calculates the absolute mean value to determine the relative importance of each variable to the overall model prediction. Local analysis traces the specific influence of features on individual predictions, providing a deeper understanding of the model's decision-making process. The application of SHAP in this study ensures the transparency and accountability of the model, as well as increasing the scientific reliability of the classification results produced by the ensemble approach.

RESULT AND DISCUSSION

Model Performance Evaluation

The classification analysis of Indonesian Liga 1 football players' market tiers was conducted using four algorithms, namely Random Forest, XGBoost, CatBoost, and Stacking

Ensemble. Performance evaluation was carried out using four main metrics, namely Accuracy, macro-f1, and f1 (high-tier), to determine the model with the best results in classifying player market values.

Table 3. Comparison of the Performance of All Classification Models

Metric Evaluation	Classification Model			
	Random Forest	**XGBoost**	**CatBoost**	**Stacking**
Accuracy	0,82	0,87	0,89	0,89
Macro-F1	0,70	0,84	0,85	0,85
F1 (High-tier)	0,50	0,79	0,78	0,78

Based on the results in Table 3, the Random Forest model achieved an accuracy of 82% with a macro-F1 value of 0.70 and an F1 (high-tier) value of 0.50. These values indicate that the model still has difficulty recognizing the High-tier class, even though its performance is quite good for the majority class. Nevertheless, Random Forest still provides stability as a baseline model, but it is not yet fully capable of effectively overcoming class imbalance. The XGBoost model showed a significant improvement in performance with an accuracy of 87%, macro-F1 of 0.84, and F1 (high-tier) of 0.79. This improvement was driven by adjustments to the scale_pos_weight parameter and threshold optimization, which made the model more sensitive to minority classes. The boosting mechanism used by XGBoost iteratively strengthens the model from previous errors, resulting in more balanced and accurate predictions for minority classes.

The CatBoost model delivers the strongest performance among all candidates, achieving an accuracy of 89%, a Macro-F1 score of 0.85, and an F1 (high-tier) of 0.78. These results indicate that CatBoost is highly effective in handling heterogeneous feature spaces and class imbalance, making it the most suitable standalone model for market value tier classification in this study. Although the Stacking Ensemble combines predictions from XGBoost and CatBoost, its performance remains identical to that of CatBoost, with an accuracy of 89%, a macro-F1 of 0.85, and F1 (high-tier) of 0.78. This outcome indicates that CatBoost already captures the dominant predictive patterns in the dataset, while the stacking layer primarily serves to validate the stability and consistency of the CatBoost predictions rather than to improve accuracy. Thus, the ensemble configuration in this study functions as a robustness-checking mechanism rather than a performance-enhancing model.

Overall, the results show that CatBoost is the best-performing model in this study, with the Stacking Ensemble confirming the reliability of its predictions rather than providing additional gains. Random Forest and XGBoost offer useful comparative baselines, but neither surpasses CatBoost in accuracy, balance, or minority-class detection. These findings reinforce that a well-regularized gradient-boosting model is sufficient to model the structure of market value tiers in Liga 1, while ensemble stacking contributes mainly to methodological validation.

Feature Importance Analysis using SHAP

Model interpretability analysis was conducted to understand the factors that most influence player market value classification. The approach used was SHAP (SHapley Additive exPlanations), which was applied to the best model, namely CatBoost. This method was used to explain the relative contribution of each feature to the model's prediction results quantitatively and visually.

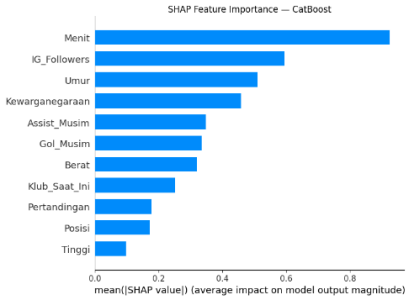

Figure 2. SHAP Feature Importance of the CatBoost Model

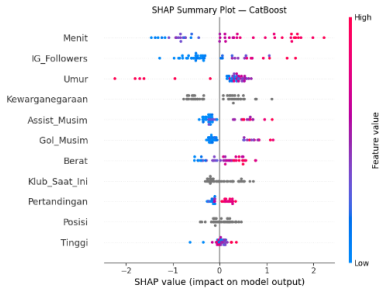

Figure 3. SHAP Summary Plot of the CatBoost Model

Two main forms of visualization were used, namely SHAP Feature Importance, which describes the average level of influence of each feature (Figure 2), and SHAP Summary Plot (Figure 3), which shows the direction and distribution of feature contributions to the model output. The SHAP visualization results show that the feature with the greatest influence on High-tier classification is Minutes Played, followed by Number of Instagram Followers (IG Followers), Age, Nationality, and Assists This Season. The Minutes Played feature has the highest positive SHAP value, indicating that the more often a player appears on the field, the greater the chance of being categorized as high-tier. Similarly, IG Followers contribute significantly, as they reflect popularity and media exposure. Conversely, features such as Height or playing position have relatively low SHAP values.

In the SHAP Summary Plot shown in Figure 3, the red color indicating high feature values is concentrated on the right side (positive SHAP values), indicating that increases in key feature values such as Minutes Played and IG Followers contribute directly to the probability of being classified as high-tier. Conversely, the blue color (low feature values) on the left side indicates a negative contribution. This pattern indicates that the model considers not only technical performance but also public exposure as an important factor. Thus, the SHAP interpretation not only reinforces the validity of the model but also confirms that performance and digital exposure dimensions have complementary contributions in determining a player's market value tiers.

Practical Implications Based on SHAP Analysis

The SHAP analysis in this study reveals that three variables minutes played, age, and social media exposure exert the most dominant influence on determining player market value tiers. These findings carry significant practical implications for Liga 1 clubs in both recruitment and squad management processes. First, the strong contribution of minutes played indicates that consistent match involvement is the most reliable indicator of market value, providing clubs with a clear basis for prioritizing players who demonstrate sustained participation and lower injury risk. Second, the negative effect of age on market value underscores the need for clubs to carefully balance peak-performance age with long-term potential when making transfer decisions. Third, the substantial role of social media exposure highlights that commercial visibility and public engagement have become important components of player valuation, enabling clubs to integrate these factors into branding strategies and non-match revenue optimization. Overall, the SHAP findings reinforce that player evaluation should not rely solely

on technical performance metrics but must also incorporate career sustainability and commercial value, both of which contribute to strengthening a club's strategic assets.

Limitations and Potential Data Bias

This study has several limitations that should be considered when interpreting the results. First, some variables were obtained from manually collected sources, such as match statistics from public platforms and market value estimates from Transfermarkt. Although commonly used in the literature, these data may contain elements of subjectivity, as market valuations are not fully objective and can be influenced by analyst opinions, market dynamics, or media exposure bias. Second, match statistics in domestic leagues often exhibit variability in quality and reporting consistency, introducing potential noise that may affect model stability. Third, several relevant variables such as player salaries, contractual clauses, and medical information are not publicly available, restricting the model to inferences based on a limited subset of features. These constraints may prevent the model from capturing the full range of determinants underlying player market value. Therefore, the findings should be interpreted with an awareness of potential data bias and limited variable completeness.

CONCLUSIONS AND RECOMMENDATIONS

This study evaluated four ensemble machine learning models (Random Forest, XGBoost, CatBoost, and Stacking Ensemble) for classifying market value tiers of Indonesian Liga 1 players in the 2024/2025 season. CatBoost demonstrated superior performance with 89% accuracy, a Macro-F1 of 0.85, an F1 (High-Tier) of 0.78, proving its reliability in modeling heterogeneous and imbalanced player data. SHAP interpretability analysis identified minutes played, age, and social media exposure as the most influential factors affecting market value, emphasizing the importance of on-field performance, career progression, and commercial visibility. These findings provide strategic implications for Liga 1 clubs in developing data-driven scouting systems, contract evaluation tools, and transfer decision-making that prioritizes players with stable performance trajectories, optimal age profiles, and measurable marketability. Tier-based classification models can mitigate overpayment risks and enhance player valuation consistency. Despite demonstrating strong predictive capability, the model's performance remains constrained by limited dataset size and potential measurement bias from manually collected variables. Future research should expand the dataset, integrate additional contextual features, explore advanced imbalance-handling strategies, and assess real-world deployment in decision-support environments.

REFERENCES

[1] J. Evans, M. Reid, And C. Watson, "Football Club Ownership And Sporting Ambition," *Managing Sport And Leisure*, Vol. 29, No. 3, Pp. 230–246, 2024.
[2] M. Varmus, "The Influence Of Domestic Players On The Success Of Football Teams," *Sage Open*, Vol. 15, No. 2, Pp. 1–13, 2025.
[3] R. Poli, L. Besson, And M. Ravenel, "Statistical Modeling Of Football Players Transfer Fees," *Int. J. Financial Studies* (Mdpi), Vol. 12, No. 1, Pp. 44–58, 2024.
[4] P. Jagielski, "Transfer Spending And Sports Performance On The Example Of Selected European Football Clubs: The Granger Causality Approach," *J. Phys. Educ. And Sport*, Vol. 22, No. 4, Pp. 1132–1140, 2022.
[5] M. Franceschi, E. López-González, And M. Roca-Cusachs, "Determinants Of Football Players Valuation: A Systematic Review," *J. Econ. Surveys*, Vol. 38, No. 1, Pp. 55–78, 2024.
[6] R. Flepp, S. Forster, And M. Stadtmann, "Outcome Bias And Informational Inefficiency In The Football Transfer Market," *Econ. Inquiry*, Vol. 62, No. 2, Pp. 512–529, 2024.
[7] I. Mchale And P. Holmes, "Modelling Transfer Fees In Professional Football," *Eur. J. Oper. Res.*, Vol. 307, No. 2, Pp. 611–624, 2023.

[8] M. Adiwiyana And S. Harymawan, "Faktor-Faktor Penentu Nilai Pasar Pemain Sepak Bola Profesional Di Indonesia," *J. Dinamika Akuntansi*, Vol. 13, No. 2, Pp. 102–112, 2021.
[9] Zulianto, A. N., Lutfillah, N. Q., And Purnomo, H., "Penentu Market Value Pemain Sepak Bola Profesional Liga 1 Indonesia Dan Thailand," Equity, Vol. 25, No. 2, Pp. 34–49, 2022.
[10] M. A. Al-Asadi And S. Taşdemir, "Predict The Value Of Football Players Using Fifa Video Game Data And Machine Learning Techniques," Ieee Access, Vol. 10, Pp. 22631–22645, 2022.
[11] D. Khalife, J. Yammine, E. Chbat, C. Zaki, And N. J. Al Maalouf, "Dynamic Financial Valuation of Football Players: A Machine Learning Approach Across Career Stages," International Journal Of Financial Studies, Vol. 13, No. 2, P. 111, 2025.
[12] A. Michael And J. Rusman, "Klasifikasi Cacat Biji Kopi Menggunakan Metode Transfer Learning Dengan Hyperparameter Tuning Gridsearch," *Jurnal Teknologi Dan Manajemen Informatika*, Vol. 9, No. 1, Pp. 37–45, 2023.
[13] H. Basri, Purnawansyah, H. Darwis, And F. Umar, "Klasifikasi Daun Herbal Menggunakan K-Nearest Neighbor Dan Convolutional Neural Network Dengan Ekstraksi Fourier Descriptor," *Jurnal Teknologi Dan Manajemen Informatika*, Vol. 9, No. 2, Pp. 79–90, 2023.
[14] R. Nahak, A. U. Bura, A. D. De Araujo, F. D. Un, B. W. Ladopurab, F. Marisa, And A. L. Maukar, "Klasifikasi Jenis Rumah Adat Malaka Menggunakan Metode Convolutional Neural Network (Cnn)," *Jurnal Teknologi Dan Manajemen Informatika*, Vol. 9, No. 2, Pp. 91–98, 2023.
[15] C. Paramita, F. A. Rafrastara, And L. I. Kencana, "Pengembangan Sistem Klasifikasi Karakteristik Siswa Berbasis Website Dengan Menggunakan Algoritma C4.5", *Jurnal Informatika: Jurnal Pengembangan It (Jpit)*, Vol. 8, No. 1, Pp. 17–25, 2023.
[16] N. Behravan And A. Razavi, "A Novel Machine Learning Method For Estimating Football Players Value In The Transfer Market," *Soft Comput.*, Vol. 25, Pp. 12265–12280, 2021.
[17] R. Sneha, P. Kumar, And S. Pradhan, "Machine Learning-Based Assessment Of Elite Football Players Based On Anthropometric And Motor Fitness Parameters With Regard To Their Playing Positions," *Sn Comput. Sci.*, Vol. 5, No. 1, Pp. 1–12, 2024.
[18] M. Sánchez García Et Al., "Factors Associated To The Market Value Of Professional Soccer Players," *Cuadernos De Psicología Del Deporte*, Vol. 22, No. 3, Pp. 148–158, 2022.
[19] K. Parpa, "Anthropometric Characteristics And Aerobic Performance Of Professional Soccer Players By Playing Position And Age," *Human Movement*, Vol. 23, No. 3, Pp. 1–10, 2022.
[20] L. Bhilawa And R. Fahriansyah, "The Influence Of Performance, Age, And Nationality On The Market Value Of Football Players," *Assets: J. Akuntansi Dan Pendidikan*, Vol. 11, No. 1, Pp. 45–54, 2022.
[21] R. Poli, R. Besson, And L. Ravenel, "Econometric Approach To Assessing The Transfer Fees And Values Of Professional Football Players," *Economies*, Vol. 10, No. 1, P. 4, 2022.
[22] A. B. Satria And D. C. Utomo, "Analysis Of Factors Affecting Market Value Of Professional Football Players In Indonesia," *Diponegoro J. Econ.*, Vol. 14, No. 2, Pp. 1–10, 2025.
[23] X. Cai, "Comprehensive Analysis Of Football Player Market Valuation: Integrating Performance Metrics And Marketability Factors," *Preprint On Researchgate*, Under Review In *Mdpi Economies*, 2024, From Https://Www.Researchgate.Net/Publication/384855068.
[24] J. Hill, B. Skinner, And T. Grosman, "A Review Of Football Player Metrics And Valuation Methods," *Journal Of Sports Analytics*, Vol. 11, No. 2, Pp. 200–218, 2025, From Https://Doi.Org/10.1080/23750472.2025.2459727.

[25] O. Müller, A. Simons, And M. Weinmann, "Beyond Crowd Judgments: Data-Driven Estimation of Market Value In Association Football," *Eur. J. Oper. Res.*, Vol. 263, No. 2, Pp. 611–624, 2017, From Https://Doi.Org/10.1016/J.Ejor.2017.05.005.
[26] S. Tamim, F. Rahman, And A. Saha, "Machine Learning-Driven Market Value Prediction For European Football Players," *J. Computational Mathematics And Data Science*, Vol. 5, No. 2, Pp. 210–226, 2025.
[27] A. Kopka And D. Fornahl, "Artificial Intelligence And Firm Growth Catch-Up Processes Of Smes Through Integrating Ai Into Their Knowledge Bases," *Small Business Economics*, Vol. 62, Pp. 63–85, 2024.
[28] S. M. Lundberg And S.-I. Lee, "A Unified Approach To Interpreting Model Predictions," *Advances In Neural Information Processing Systems*, Vol. 30, Pp. 4765–4774, 2017, From Https://Doi.Org/10.48550/Arxiv.1705.07874.
[29] C. Molnar, *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable* (2nd Ed.), 2022.