International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Machine Learning-Based Predictive Framework for Patent Infringement Detection: Enhancing Intellectual Property Protection Through Hybrid Ensemble Models

Kang Jin Gang[a]*, Ang Ling Weay[b]

[a,b]*Malaysia University of Science and Technology (MUST), Block B, Encorp Strand Garden Office, No. 12, Jalan PJU 5/5, Kota Damansara, 47810 Petaling Jaya, Selangor, Malaysia*

[a]*Email: kang.jingang@phd.must.edu.my*

[b]*Email: dr.ang@must.edu.my*

Abstract

Patent infringement poses significant risks to innovation and economic growth. Traditional intellectual property (IP) protection methods are often reactive, expensive, and inefficient for large-scale patent management. This study introduces an optimized machine learning framework designed to predict patent infringements proactively. The research evaluates the performance of Random Forest, Support Vector Machines (SVM), and Logistic Regression on a curated dataset enriched with patent citations, legal status, and family size. The study employs Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and Recursive Feature Elimination (RFE) for feature selection. A novel hybrid ensemble model integrating Random Forest and SVM is developed, achieving 75% precision, 95% recall, and an F1-score of 84%, outperforming baseline models. The findings contribute to IP management by offering a scalable predictive framework that minimizes litigation costs and enhances proactive infringement detection.

Keywords: Patent Infringement Prediction; Machine Learning; Hybrid Ensemble Algorithm; Intellectual Property Management; Data Balancing; Feature Selection.

Received: 2/27/2025
Accepted: 4/25/2025
Published: 5/5/2025

---

* Corresponding author.

1. Introduction

Intellectual property (IP) protection plays a critical role in driving technological innovation and sustaining economic growth. As patents form a central mechanism through which inventors and organizations protect their innovations, the integrity of patent systems directly influences research incentives and market competitiveness. However, the exponential rise in patent filings globally has introduced significant challenges. According to the World Intellectual Property Organization (WIPO), over 3.4 million patent applications were filed globally in 2021, with Asia accounting for more than 67% of this volume [1] (WIPO, 2022). This rapid growth has been accompanied by a corresponding surge in patent infringement cases, making it increasingly difficult for institutions and enterprises to manage and protect their intellectual assets effectively.

Traditional methods of patent protection, which primarily involve manual patent analysis, post-violation enforcement, and prolonged legal disputes, are no longer adequate in the current digital and data-intensive environment. These reactive approaches are resource-intensive, time-consuming, and often fail to provide timely intervention against potential threats. Moreover, the legal complexities involved in patent litigation often lead to high costs, making enforcement unfeasible, especially for small and medium enterprises (SMEs) and academic institutions [2] (Zhao and his colleagues 2020).

In response to these limitations, machine learning (ML) has emerged as a transformative approach capable of revolutionizing intellectual property management. ML algorithms are designed to process vast datasets, uncover complex patterns, and generate predictive insights—capabilities that are well-suited to the dynamic and multifaceted nature of patent data [3] (Nguyen and his colleagues 2021). Through the integration of predictive analytics, organizations can shift from reactive to proactive IP protection strategies, enabling them to detect potential infringement risks early and act accordingly.

Recent studies have demonstrated the effectiveness of ML models such as Random Forest, Support Vector Machines (SVM), and Logistic Regression in classifying patent relevance, forecasting litigation likelihood, and detecting infringement indicators [4,5] (Son and his colleagues 2022; Lee and his colleagues 2022). These techniques offer substantial improvements over manual review processes by increasing the speed, accuracy, and scalability of infringement detection. Furthermore, by incorporating feature engineering and data preprocessing techniques, ML models can achieve high levels of precision and recall, which are critical for minimizing false negatives and ensuring comprehensive protection [6] (Juranek & Otneim, 2021).

This study aims to advance current research by developing an optimized ML-based framework for predicting patent infringements. It evaluates the comparative performance of different algorithms and introduces a hybrid ensemble model that leverages the strengths of both Random Forest and SVM. By incorporating data balancing methods such as Synthetic Minority Oversampling Technique (SMOTE) and feature selection techniques like Recursive Feature Elimination (RFE), the proposed framework addresses common challenges such as class imbalance and model overfitting. The goal is to offer a robust, scalable, and accurate predictive system that enhances IP management, reduces litigation costs, and supports informed decision-making across diverse organizational contexts.

2. Methodology

Recent advancements in machine learning (ML) have opened up promising avenues for addressing the complex challenge of patent infringement detection. Traditional intellectual property (IP) protection strategies, which rely heavily on post-hoc legal enforcement and manual patent monitoring, are increasingly inadequate in the face of rising patent volumes and sophisticated infringement tactics. In contrast, ML algorithms can process large-scale and high-dimensional datasets to uncover intricate patterns that may indicate potential infringement. This proactive approach enables earlier intervention, significantly reducing enforcement costs and enabling more strategic and informed IP management decisions. By leveraging these capabilities, this study aims to develop a robust ML framework for predicting patent infringements, focusing on the integration of multiple algorithms to enhance predictive performance and reliability.

The methodological foundation of this research involves a structured process that includes data collection, preprocessing, model selection, hybrid model development, and thorough evaluation. The dataset used in this study was curated from publicly available patent databases and included essential features such as forward and backward citations, legal status (e.g., granted, pending, or revoked), and patent family size. These attributes were chosen for their strong relevance in previous studies linking them to infringement risk, as citation frequency and legal status often reflect a patent's influence and enforceability [7,8] (Qi, 2014; Cremers, 2004).

To ensure data quality and model effectiveness, several preprocessing steps were applied. Missing values were addressed using mean or mode imputation techniques, which are well-suited for preserving data integrity without introducing bias in cases of sparse missingness. Numerical features were normalized to bring all values into a comparable range, improving the performance of algorithms sensitive to scale, such as Support Vector Machines (SVM). Feature engineering was also conducted to derive new variables that could offer deeper insights—such as citation ratios or patent age—thereby enriching the input space for the models. Label encoding was used to convert categorical variables into numeric formats, making the dataset compatible with ML algorithms that require numerical input.

For the algorithmic core of the study, three widely adopted ML models were selected to establish baseline performance comparisons: Random Forest (RF), Support Vector Machines (SVM), and Logistic Regression (LR). Random Forest was chosen due to its ensemble nature, which mitigates overfitting by averaging multiple decision trees, thereby improving generalizability and robustness in complex datasets [9] (Breiman, 2001). SVM was selected for its strength in handling high-dimensional data and its ability to construct optimal decision boundaries in binary classification problems like infringement detection [10] (Cortes & Vapnik, 1995). Logistic Regression, while simpler, was included for its interpretability and effectiveness in capturing linear relationships, serving as a useful benchmark to gauge the added value of more complex models.

Each model was evaluated based on four key performance metrics: precision, recall, F1-score, and the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC). These metrics were chosen to provide a comprehensive assessment of each model's effectiveness, particularly given the class imbalance typical in infringement datasets. Precision measures the model's accuracy in identifying true infringement cases, recall

assesses the ability to capture all actual infringement instances, F1-score balances the trade-off between precision and recall, and ROC-AUC provides an aggregate measure of performance across all classification thresholds. This holistic evaluation framework ensures that the selected model not only performs well statistically but is also practically effective for real-world deployment in proactive IP management systems.

By applying this rigorous methodology, the study seeks to build a predictive framework that not only advances theoretical knowledge but also provides a practical tool for enhancing the protection of intellectual property through modern machine learning techniques.

3. Results and Analysis

To improve predictive performance in patent infringement detection, a hybrid ensemble model was developed by integrating two of the best-performing base classifiers: Random Forest (RF) and Support Vector Machines (SVM). The motivation for combining these algorithms lies in their complementary strengths. Random Forest is an ensemble of decision trees known for its robustness to noise and ability to capture non-linear feature interactions. In contrast, SVM is highly effective in high-dimensional spaces and excels at creating precise decision boundaries. By integrating these two models, the ensemble leverages RF's capacity to handle diverse feature distributions and SVM's proficiency in separating complex data points. A weighted voting mechanism was employed in which the final prediction was based on the confidence scores assigned by each model. This strategy ensures that the model with higher predictive confidence for a given instance has a proportionally greater influence on the final decision.

To further refine the hybrid model, several enhancement techniques were integrated. First, data balancing was addressed using the Synthetic Minority Oversampling Technique (SMOTE). This step was crucial because the original dataset exhibited a class imbalance typical of infringement scenarios, where positive (infringement) cases are significantly fewer than negative ones. Without addressing this imbalance, the model risked biasing toward the majority class, thereby reducing its ability to detect true infringement cases. SMOTE effectively mitigated this by generating synthetic samples for the minority class, enabling the model to learn more balanced decision boundaries.

Second, Recursive Feature Elimination (RFE) was applied to identify and retain only the most relevant features for prediction. Patent datasets often include high-dimensional data, which may contain irrelevant or redundant features. These can lead to overfitting and degrade model performance. RFE systematically removes less important features and retains those contributing most to prediction accuracy, thus reducing dimensionality, improving model generalization, and decreasing computational load.

Third, hyperparameter tuning was carried out using grid search in combination with cross-validation. Grid search exhaustively searches through a manually specified subset of hyperparameter values, while cross-validation ensures the model's stability and robustness across different data splits. This process was critical for optimizing each base learner (RF and SVM) and the final ensemble, ensuring that the model achieved optimal performance rather than relying on default or arbitrary parameters.

The experimental evaluation was conducted using a 70/30 train-test split. Table 1 presents the performance metrics for each model:

Table 1: Analysis Results

Model	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	61%	72%	66%	0.74
Random Forest	68%	86%	76%	0.85
SVM	71%	88%	78%	0.88
Hybrid Model	**75%**	**95%**	**84%**	**0.91**

The hybrid ensemble model outperformed all individual classifiers across every metric. Notably, the model achieved a precision of 75%, indicating its ability to correctly identify a large proportion of actual infringement cases among all those predicted as infringements. This is essential in real-world applications where false positives can lead to unnecessary legal scrutiny and resource expenditure. Its recall of 95% reflects an outstanding ability to detect actual infringement cases, a critical feature in minimizing undetected violations. The F1-score of 84% demonstrates a strong balance between precision and recall, further validating the model's suitability for practical deployment. Finally, the ROC-AUC of 0.91 shows excellent discriminatory power, with the model being able to distinguish between infringement and non-infringement cases more effectively than the individual models.

These results justify the integration of ensemble learning and data preprocessing techniques in building an advanced predictive model. The hybrid model not only addresses the technical challenges of class imbalance and feature redundancy but also provides a scalable, high-performance solution for patent infringement detection. This validates the core hypothesis of the study: that combining the strengths of diverse ML algorithms within a well-engineered framework yields significantly superior results compared to standalone models.

4. Discussion

The empirical results from this study clearly illustrate the substantial benefits of utilizing ensemble learning methods in predicting patent infringement. By combining the predictive capabilities of Random Forest (RF) and Support Vector Machines (SVM), the hybrid ensemble model outperformed each of the individual models—including Logistic Regression, RF, and SVM—across all major performance metrics. These metrics included precision, recall, F1-score, and the area under the ROC curve (ROC-AUC). The superior performance of the hybrid model highlights the synergistic effect of blending two powerful algorithms. While Random Forest excels at managing noisy or imbalanced data through its bagging approach and decision tree ensembles, SVM is highly effective at constructing optimal decision boundaries in high-dimensional feature spaces. Their integration not only increased the accuracy of the model but also improved its overall stability and robustness, making it well-suited for the complexities of real-world patent infringement detection.

One of the key technical interventions that significantly enhanced model performance was the application of the Synthetic Minority Oversampling Technique (SMOTE). In the domain of patent infringement, datasets are typically skewed, with far fewer infringement cases compared to non-infringement ones. This imbalance can lead to model bias, where the learning algorithm disproportionately favors the majority class and fails to recognize the critical minority class—in this case, actual infringements. SMOTE addressed this by generating synthetic samples of the minority class, effectively balancing the dataset and allowing the model to learn a more accurate representation of both classes. The outcome was a notable increase in recall, indicating that the model became significantly better at detecting true positives. This capability is vital for proactive intellectual property (IP) protection, where missing an infringement case could lead to costly legal consequences and compromised innovation assets.

Additionally, Recursive Feature Elimination (RFE) played a crucial role in refining the model by eliminating non-informative or weakly correlated features. Patent datasets often contain numerous attributes such as citation counts, legal status, and patent family size, many of which may overlap or contribute little to predictive power. Without feature selection, the model risks overfitting, longer training times, and reduced interpretability. By employing RFE, the study ensured that only the most relevant features were retained, which improved both computational efficiency and model interpretability. This enhancement was particularly important in reducing dimensionality and focusing the model on factors most indicative of infringement risk.

Furthermore, the study implemented iterative optimization strategies involving grid search and cross-validation to fine-tune the model's hyperparameters. This process systematically explored different parameter settings and assessed their performance using multiple data partitions, ensuring that the model did not simply memorize training data but instead learned generalizable patterns. These techniques contributed to the model's ability to maintain high performance when exposed to new, unseen data—a critical requirement for any predictive system intended for practical use in legal and corporate settings.

Despite the promising outcomes, several limitations of the study must be acknowledged. First, although the dataset included important features such as citations and legal status, it lacked deeper semantic information from the patent text itself. Many infringement cases hinge on nuanced differences in claim language or technological specifications that cannot be captured through structured metadata alone. Additionally, the dataset's applicability may be limited by jurisdictional variation in how patents are written, classified, and enforced. The legal frameworks, language standards, and documentation practices differ globally, which could hinder the model's generalizability beyond the original dataset's context.

Another practical constraint lies in the computational resources required to train and deploy the hybrid ensemble model. Techniques such as SMOTE, RFE, and cross-validation are computationally intensive, especially when applied to large datasets. This may pose scalability issues or limit adoption by smaller organizations with limited technological infrastructure.

To address these challenges and extend the utility of the proposed framework, several future directions are recommended. One important avenue is the integration of Natural Language Processing (NLP) to extract and

analyze the full textual content of patents. Semantic models like BERT or GPT could uncover subtler dimensions of patent similarity and infringement risk, improving prediction depth and precision. Expanding the feature space to include variables such as examiner notes, prosecution history, or jurisdiction-specific attributes could also provide richer training data and allow for more nuanced modeling. Additionally, developing domain-specific models tailored to particular industries—such as biotechnology, information technology, or mechanical engineering—may enhance accuracy, as infringement patterns often vary by sector.

Future work should also focus on real-time integration of the model into existing IP management platforms. A streamlined, computationally efficient version of the hybrid framework could assist patent attorneys, corporate IP teams, and legal tech providers in assessing risk during patent application reviews or portfolio management activities. Moreover, validating the model on international datasets would provide crucial insights into its cross-jurisdictional effectiveness, an important step for global implementation.

Lastly, as machine learning tools begin influencing legal strategies and decisions, ethical and legal considerations must be foregrounded. It is essential that such systems are transparent, interpretable, and free from bias, especially when their outputs may affect litigation or investment decisions. Ensuring compliance with data privacy and intellectual property laws will also be vital in maintaining trust and legality in real-world deployments.

In conclusion, the study provides compelling evidence that a hybrid machine learning framework, enhanced by thoughtful data processing and optimization techniques, offers a powerful tool for proactive patent infringement detection. While the model shows strong potential, further refinements—such as NLP integration, jurisdictional adaptation, and ethical oversight—are necessary to enable widespread and responsible adoption in dynamic global IP landscapes.

5. Conclusion

This study introduces a machine learning-based predictive framework aimed at improving the detection of patent infringement, offering a more proactive and intelligent approach to intellectual property (IP) management. Central to this framework is the integration of two powerful classification algorithms—Random Forest and Support Vector Machines (SVM)—into a hybrid ensemble model. This model is further enhanced by the application of advanced machine learning techniques, specifically the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and Recursive Feature Elimination (RFE) to optimize feature selection. Together, these components contribute to a system that not only demonstrates high accuracy and precision but also maintains robustness across diverse and imbalanced datasets.

Unlike traditional IP protection methods, which often rely on reactive litigation and manual monitoring, the proposed framework offers a scalable, data-driven alternative that can detect infringement risks before they escalate. This predictive capability is particularly valuable in today’s innovation-driven economy, where organizations manage vast portfolios of patents and face increasingly complex infringement threats. The model’s ability to integrate structured patent metadata and identify patterns of infringement enhances both

strategic decision-making and operational efficiency in IP management.

While the framework developed in this study represents a significant advancement, it also lays the groundwork for future exploration. One promising direction is the incorporation of deep learning techniques, such as convolutional neural networks (CNNs), which could be used to analyze the semantic content of patent texts, including claims and descriptions. Such models have the potential to capture subtle textual nuances and similarities that structured data may overlook. Another opportunity lies in the application of graph-based algorithms to map and analyze patent citation networks. By modeling the relationships between patents, these approaches could uncover hidden linkages and prior art, offering deeper insights into potential infringement risks.

Furthermore, the conceptual foundation established here can be extended to other domains of intellectual property, including trademarks and copyrights. These areas, while distinct in legal structure and enforcement, share common challenges related to infringement detection and portfolio management. Adapting the machine learning framework to accommodate different types of IP data would broaden its utility and support a more comprehensive approach to digital IP protection.

Ultimately, this research bridges the gap between theoretical innovation in machine learning and practical application in intellectual property management. It not only validates the effectiveness of combining ensemble models with intelligent data preprocessing techniques but also demonstrates the feasibility of deploying such systems in real-world settings. By advancing from reactive enforcement to predictive protection, the proposed framework marks a critical step toward a more efficient, scalable, and intelligent future for data-driven IP governance.

References

[1] World Intellectual Property Organization (WIPO), World Intellectual Property Indicators 2022. Geneva, Switzerland: WIPO, 2022. [Online]. Available: https://www.wipo.int/publications/en/details.jsp?id=4589

[2] Z. Zhao, L. Wang, and X. Liang, “Patent litigation risk analysis for SMEs based on machine learning,” Technovation, vol. 94–95, p. 102089, 2020, doi: 10.1016/j.technovation.2020.102089.

[3] H. D. Nguyen, N. H. Tran, M. T. Nguyen, and T. Q. Dinh, “Artificial intelligence in intellectual property management: Applications and research challenges,” IEEE Access, vol. 9, pp. 123154–123172, 2021, doi: 10.1109/ACCESS.2021.3110103.

[4] J. Son, H. Lim, and J. Lee, “Deep learning-based infringement risk prediction using patent documents,” Journal of Informetrics, vol. 16, no. 1, p. 101203, 2022, doi: 10.1016/j.joi.2021.101203.

[5] S. Lee, H. Park, and H. Kim, “Patent infringement prediction using machine learning techniques: Evidence from U.S. patents,” Technological Forecasting and Social Change, vol. 174, p. 121253, 2022,

doi: 10.1016/j.techfore.2021.121253.

[6] S. Juranek and H. Otneim, “Predicting patent litigation with machine learning,” Research Policy, vol. 50, no. 2, p. 104154, 2021, doi: 10.1016/j.respol.2020.104154.

[7] Y. Qi, “Patent characteristics and patent litigation: Empirical evidence from China,” J. World Intellect. Prop., vol. 17, no. 5–6, pp. 204–217, 2014, doi: 10.1111/jwip.12035.

[8] K. Cremers, “Determinants of patent litigation in Germany,” Centre for European Economic Research, ZEW Discussion Paper No. 04-072, 2004. [Online]. Available: https://ftp.zew.de/pub/zew-docs/dp/dp04072.pdf

[9] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[10] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.