

A Novel Approach to Predict European Mergers Using Machine Learning Algorithm

Pankaj Gupta^{1*}, Dr. Satyen M. Parikh², Dr. Meghna B. Patel³

^{1*}FCA, Ganpat University, Gujarat, India, pangup_74@yahoo.com

²FCA, Ganpat University, Gujarat, India, satyen.parikh@ganpatuniversity.ac.in

³FCA, Ganpat University, Gujarat, India, meghna.patel@ganpatuniversity.ac.in

Abstract

The European Commission (EC) plays a central role in reviewing mergers and acquisitions (M&A) within the European Economic Area to safeguard market competition and protect consumer interests. The decision-making process is complex, involving assessments of market dominance, competitive overlap, and industry-specific factors, often leading to delayed outcomes for firms and stakeholders. This study applies machine learning (ML) techniques to develop a predictive framework for anticipating EC merger decisions. We evaluate three established classifiers—Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB) on a structured dataset of past EC M&A cases.

To enhance predictive accuracy, a hybrid model using a stacking ensemble approach is proposed, where outputs from base models are combined using a Logistic Regression meta-learner. The model is assessed using standard classification metrics, including accuracy, F1 Score, AUC-ROC, precision, and recall. Empirical results demonstrate that the hybrid model achieves improved performance over individual classifiers, particularly in recall and F1 Score two metrics that are critical for imbalanced regulatory datasets. This research contributes a data-driven approach to merger control analysis and offers practical implications for policymakers, legal analysts, and corporate advisors navigating EC merger regulations.

Keywords : Merger control, ensemble learning, European Commission, machine learning, regulatory prediction

Introduction

Mergers and acquisitions (M&A) represent a critical aspect of corporate strategy, allowing firms to grow, enter new markets, consolidate operations, and gain access to novel technologies or competencies. In globalized and highly competitive economies, M&A transactions often influence entire industries by reshaping market structures and triggering shifts in pricing, innovation, and consumer welfare. As such, regulatory oversight plays a vital role in maintaining fair competition and ensuring that corporate consolidation does not lead to monopolistic dominance or consumer harm [1]. Within the European Union (EU), the European Commission (EC) is the primary authority responsible for merger control. The EC's assessment of proposed transactions is governed by the EU Merger Regulation, which mandates a thorough analysis of whether a merger would significantly impede effective competition, particularly through the creation or strengthening of a dominant position. This involves a case-by-case review of market share data, industry characteristics, horizontal and vertical overlaps, and potential anti-competitive effects. While essential to market regulation, this process is often resource-intensive and time-consuming, introducing delays and uncertainty for merging firms, legal advisors, and policymakers [2][4].

Traditional merger assessments rely on qualitative economic reasoning, legal precedent, and internal expertise, which—although robust—are not immune to cognitive bias, subjectivity, or inconsistency across cases. In response to these challenges, there is growing interest in integrating data-driven

approaches to support decision-making in merger control. Machine learning (ML), in particular, has demonstrated strong potential in domains involving classification, pattern recognition, and regulatory forecasting. Its ability to extract insights from large historical datasets and adapt to complex, non-linear relationships makes it well-suited to supplement traditional regulatory tools [4]. As illustrated in Figure 1, historical merger cases were evaluated for training and testing the predictive model [3].

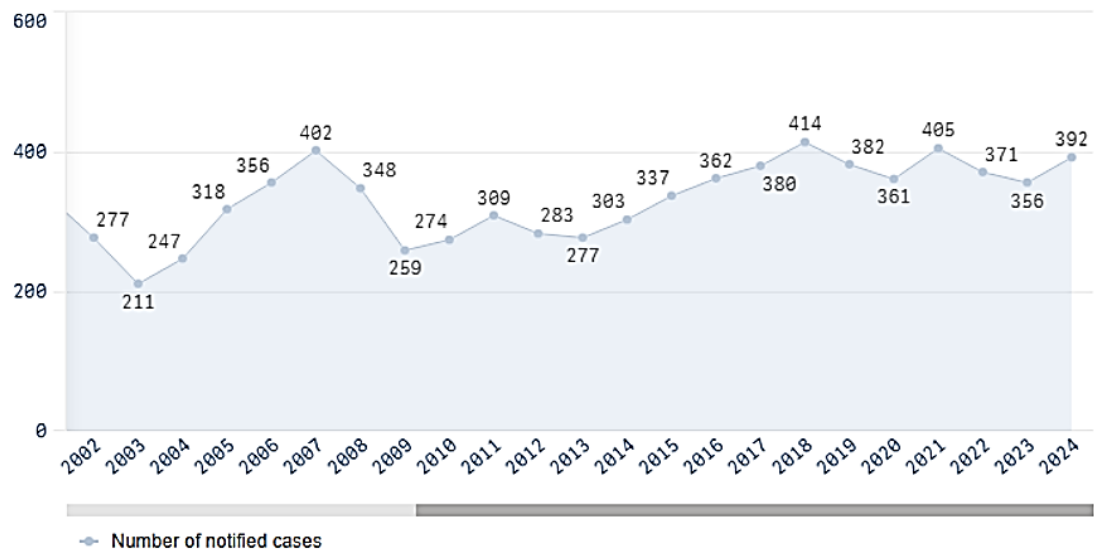


Figure 1
Evaluation of Historical Merger Cases for Predictive Modeling

Note. This figure illustrates the assessment process and structure of merger cases used for machine learning model development. Adapted from European Commission data [3].

Despite the emergence of ML in legal analytics, relatively few studies have focused specifically on applying ML techniques to predict merger outcomes under EU competition law. Even fewer have explored ensemble or hybrid approaches that combine the predictive strengths of multiple classifiers. This research addresses that gap by proposing a hybrid machine learning model that integrates Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB) classifiers through a stacking ensemble. The goal is to predict EC merger decision outcomes such as unconditional approval, conditional approval, or in-depth investigation with improved accuracy, recall, and interpretability.

Using a structured dataset compiled from publicly available EC merger cases, the study compares the performance of individual classifiers and the proposed hybrid model using key evaluation metrics, including F1 Score, AUC-ROC, precision, and recall. The results show that the hybrid model outperforms the standalone models, particularly in capturing rare but important outcomes like blocked or conditionally cleared mergers[1].

This paper contributes to the intersection of regulatory technology, machine learning, and competition policy. It provides a methodological framework for applying ensemble learning to legal decision modeling and offers practical implications for firms and regulators seeking to anticipate merger outcomes with greater certainty and transparency.

Literature Review

The increasing complexity of merger regulation, especially within the European Union (EU), has sparked scholarly interest in leveraging machine learning (ML) to enhance legal and economic decision-making. This literature review synthesizes prior work across four major themes: predictive modeling of M&A activity, legal text classification, ensemble learning for regulatory outcomes, and antitrust decision modeling.

1. Predictive Modeling of M&A Activity

It explored how machine learning can detect early signals of merger activity by analysing unstructured corporate disclosures. Using LASSO-based logistic regression, the study examined textual data from U.S. firms' 10-K filings to identify narrative elements associated with future M&A transactions. The analysis demonstrated that certain language patterns in corporate reports serve as reliable predictors of impending mergers. This work supports the broader argument that regulatory and strategic behaviours can be forecasted using supervised learning on textual data sources, which are often underutilized in traditional econometric analysis[6]. Similarly, The combined textual features with financial variables to forecast M&A targets and acquirers. Their text regression framework proved that linguistic patterns particularly those referencing firm weakness, tax structures, or strategic synergies are strong predictors when combined with conventional metrics [7].

The approached prediction through product-market similarity by using cosine similarity on TF-IDF-weighted 10-K product descriptions. Their “text-based network industry classification” outperformed traditional SIC codes in modeling competitive overlap, providing a more nuanced basis for detecting potential merger synergies or antitrust risks. However, these models primarily focused on firm behavior and lacked direct application to regulatory outcomes [8].

2. Legal Text Classification and Judicial Prediction

Several studies have demonstrated that legal decision-making often seen as opaque can be forecasted using ML techniques. SVM and CNN models to predict outcomes at the European Court of Human Rights by combining factual descriptions and legal reasoning [9][10]. It extended this to user-generated legal text, proving that even layperson-submitted narratives can yield accurate predictions through NLP [9][10].

It applied deep learning (LSTM, CNN, MLP) to legal document classification using semantically enriched representations. Their findings showed improved accuracy compared to traditional classifiers, although interpretability remained limited [11]. In a large-scale judicial forecasting study, a time-evolving Random Forest model was introduced to predict U.S. Supreme Court votes and decisions. Their results, while robust, emphasized the need for high quality, structured historical data, an area often lacking in EU merger enforcement documentation [12].

3. Ensemble Learning in Regulatory Compliance

Ensemble learning methods have shown promise in domains where prediction errors carry high policy or financial risk. It assessed how ensemble learning can enhance predictive accuracy in domains with high regulatory and compliance stakes. Their research focused on stacking techniques that integrate multiple classifiers such as decision trees, SVMs, and boosting algorithms through a meta-learning layer. The results showed that such hybrid models consistently outperform single classifiers, particularly in imbalanced datasets where classification errors carry significant consequences. Their findings are especially relevant to legal and regulatory prediction problems, where interpretability and precision are critical [13].

It applied NLP and supervised learning to EC merger decisions, extracting narrative patterns linked to anticompetitive risks. The study confirmed that textual features from legal documents can enhance

merger outcome forecasting. However, it did not implement ensemble models, limiting its predictive performance and generalizability [14].

4. Economic and Policy-Oriented Antitrust Analysis

Early econometric analyses evaluated 96 EC merger cases and concluded that high market share, significant entry barriers, and collusion risks were strong predictors of conditional approval or prohibition. Their work laid the empirical foundation for identifying economic variables relevant to regulatory decisions [15].

It examined the relationship between antitrust policy enforcement and innovation by studying a regulatory shift that exempted many smaller horizontal mergers from review. Using a novel NLP-based patent similarity method, the study detected mergers between close competitors and found that those not reviewed by regulators showed a significant decline in subsequent patent activity. The findings support a deterrence theory of merger control where stronger regulatory scrutiny incentivizes firms to maintain innovation. Additionally, the research presents a theoretical model illustrating how firms adjust merger behaviour in response to enforcement probability. Beyond innovation effects, the study also highlights increased market concentration and reduced labour share in non-reviewed mergers, reinforcing the broader socioeconomic consequences of antitrust policy [16].

Synthesis

Collectively, the literature underscores the viability of machine learning for merger analysis whether predicting firm behavior, legal decisions, or regulatory risk. However, few studies have applied stacked ensemble learning to predict outcomes under the EU Merger Regulation, using actual EC decision texts. This research addresses that gap by combining Random Forest, Support Vector Machine, and Naive Bayes classifiers in a stacking architecture, enhancing predictive accuracy and interpretability. In doing so, it contributes a data-driven framework to support legal analysts, policymakers, and corporate strategists navigating European merger control.

Proposed Research Model

In this study, we introduce a hybrid machine learning model to predict European Commission (EC) decisions on merger and acquisition (M&A) cases. The model is based on a **stacked ensemble learning approach**, which combines the outputs of multiple base classifiers to improve overall predictive performance. The motivation behind this hybrid model is to capture the diverse decision patterns present in complex regulatory data, which individual classifiers may not fully model alone.

The below figure 2 shows large set of competition cases in the form of PDF documents were collected from the European Commission's website. Each PDF document contains the full text of the decision along with metadata, such as the number of pages and the length of the document etc.

For Data Preprocessing, Text Cleaning and Missing Data was handled. Before feature extraction, the text was cleaned by removing stop words, numbers, and special characters. This ensured that only meaningful information was retained for the model. Missing values (replaced to 0) in the metadata features were handled using a Simple Imputer.

The hybrid model utilizes three supervised machine learning classifiers as base learners.

1. **Random Forest (RF)**: an ensemble of decision trees that handles non-linearity and reduces overfitting via bagging.
2. **Support Vector Machine (SVM)**: effective in high-dimensional spaces and known for its strong theoretical guarantees.
3. **Naive Bayes (NB)**: a probabilistic classifier that performs well with categorical data and is computationally efficient.

Each of these models is trained independently on the same pre-processed dataset. During the prediction phase, each base model generates an output either a predicted class or a probability score. These outputs are then combined and passed as input features to a **meta-learner**.

The **meta-learner** is trained on the predictions of the base classifiers using a separate validation fold to avoid data leakage. In our primary setup, we use **Logistic Regression** for its simplicity and interpretability. We also evaluate **Gradient Boosting** as an alternative meta-learner to capture complex non-linear interactions.

To ensure robustness, we adopt **n-fold cross-validation** for both training the base models and the meta-learner. This two-level training process helps avoid overfitting and allows each model to generalize better to unseen data. Hyperparameters for all classifiers are tuned using grid search.

This stacked model is expected to outperform individual classifiers by exploiting their complementary strengths. For instance, while RF may handle complex feature interactions, it may be less precise in some linear boundaries where SVM excels. NB, though less accurate on its own, can contribute useful probabilistic patterns that improve the final meta-prediction.

In the regulatory domain of EC merger control, where decisions are influenced by multifactorial and often impervious factors, this hybrid approach offers a more nuanced and resilient prediction mechanism.

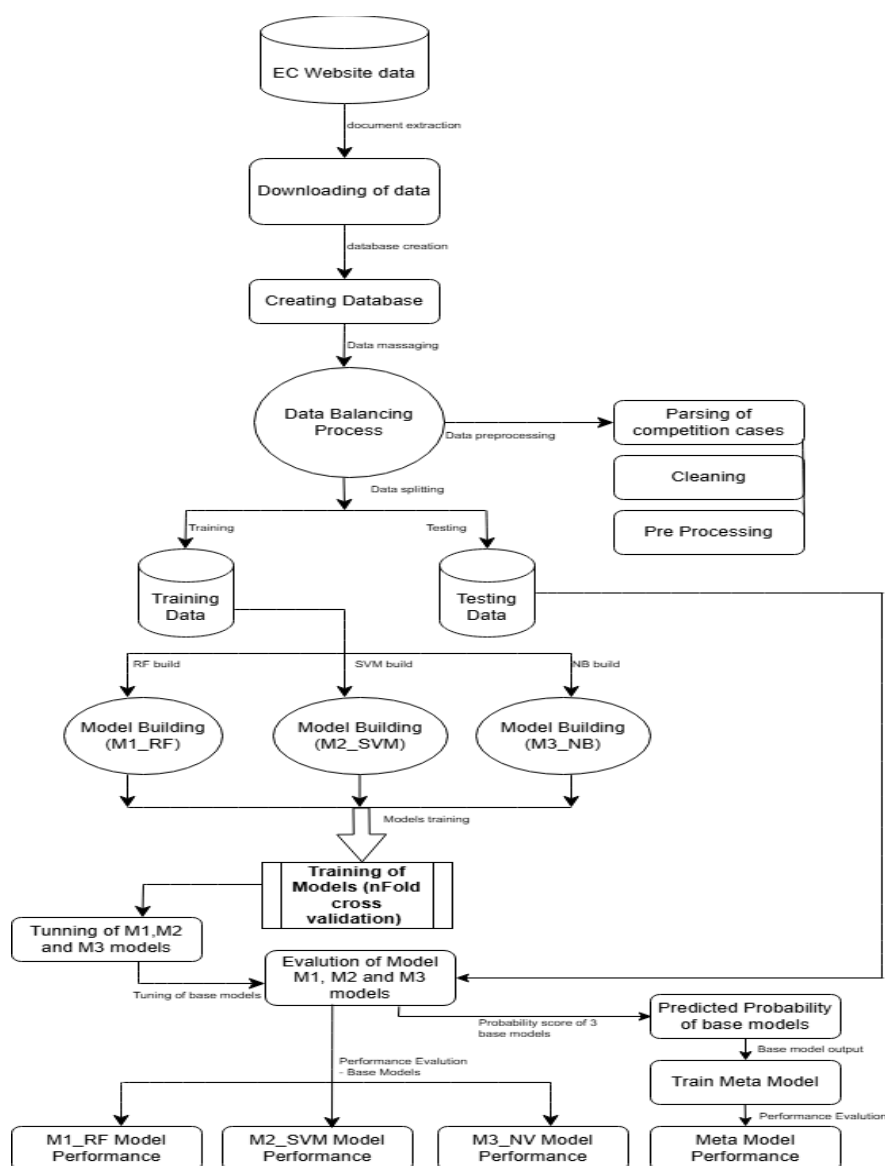


Figure 2

Proposed Hybrid Machine Learning Model for Predicting EC Merger Decisions

Note. The model combines Random Forest, SVM, and Naive Bayes classifiers through a stacked ensemble using logistic regression as a meta-learner. Data were pre-processed from official EC decision documents.

Experiment Results

Dataset

This study examined competition cases reviewed under the EU Merger Control Regulation 2004, with notification dates between 1 May 2004 and 20 October 2024. Merger related decisions documents filed by corporates intended to go with Merger and Acquisition agreement, can be accessible through the European commission competition search tool. Manually downloading of each decision documents in PDF format over the nearly 20-year period would be an arduous task. To streamline this process, a program was created to automatically download the documents from the EC search tool's website as shown in the below Figure 3 .

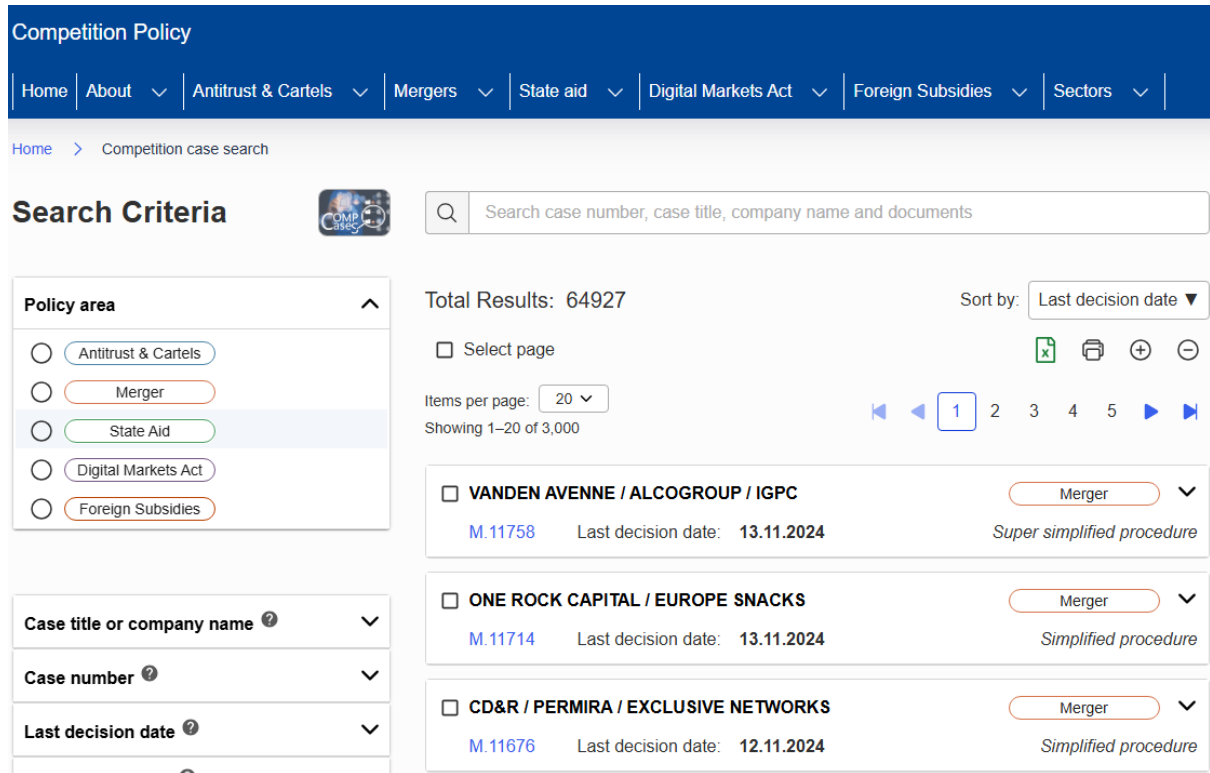


Figure 3
European Website for Downloading EC Merger Decision Document

Note. A custom utility was developed to batch-download and preprocess merger-related documents from the European Commission competition database.

Initially, it was noted that the EU merger website searched 64927 documents, which were filed in the form of Antitrust & cartels, Mergers and foreign subsidiaries. The utility was restricted to download only mergers related key documents. Total of 6920 unique merger cases as of 20 October 2024 were downloaded in the batches. Other information obtained from the webpage were the case number, title, notification date, and NACE or economic activity concerned.

In terms of decision extensiveness, data was gathered in all situations where a legal decision document was available. This covers all cases resolved in the first part of a review (Art. 6(1)(a), 6(1)(b), 6(1)(c), and 6(2)), as well as all cases determined in the second phase of an investigation (Art. 8(1), 8(2), and 8(3)). It should be noted that this covers all matters settled via a 'simplified method', provided that a legal decision document exists.

The initial distribution by article shows that the majority of cases (4364) were approved unconditionally under Simplified Procedure, with a standardized decision of only three pages long. Such cases were excluded in this study as they only contain short descriptions of the merging parties' business activities. Some more cases were removed, and these were found to be duplicate. While this step diminished the dataset to 2221 cases for parsing and pre-processing, this made the case structure consistent by removing the noise attributed to how a merger decision was written. These merger cases are in multiple languages. However, we restricted the downloader utility to download English language documents only. Table 1 presents a summary of competition case statistics available on the European Commission's merger database [17].

Table 1
Competition Case Statistics from the European Commission Website

Description	Competition Cases
Total Documents available	64927
Merger related Document	9670
Normal Procedure	4380
Ongoing Merger cases	23

Note. Data retrieved from the European Commission (2024)

European Commission Merger Document Structure:

The European Commission assesses mergers under a structured, two-phase regulatory process guided by the EU Merger Regulation. In the initial stage, known as **Phase One**, the Commission conducts a preliminary review to determine whether the transaction raises significant competition concerns. If no major issues are identified, the merger is typically approved within 25 working days. However, when a proposed transaction appears to threaten market competition, particularly in cases involving high market concentration or potential dominance, undergoes a second-phase review under Article 8 of the EU Merger Regulation, which allows for additional data gathering and stakeholder consultation (European Commission, 2023). This stage allows the Commission additional time typically 90 working days to thoroughly analyse the potential effects on the internal market, gather further data, and consult third parties. Depending on the outcome, the Commission may approve the merger unconditionally, impose structural or behavioural remedies, or prohibit the transaction altogether. This multi-stage approach ensures that merger control decisions are both timely and evidence-based, aiming to balance corporate integration with the need to maintain fair competition across the European Union. The standard structure (from Article 6(1)(b)/ of the Merger Regulation) of European Commission merger decisions is outlined in Table 2 [19].

Table 2
Standard Structure of European Commission Merger Decision Documents

S. No.	Section Name	Section Description
1	HEADER INFORMATION	Case number, date, parties, and transaction details
2	INTRODUCTION	Overview, legal basis, and jurisdiction
3	THE PARTIES	Detailed information about the merging firms
4	THE OPERATION	The structure of the deal
5	DIMENSION	Defining the relevant product and geographic markets.
6	COMPETITIVE ASSESSMENT	This section addresses about the analysis of competitive effects. This includes horizontal, vertical and conglomerate effects as well.
7	REMEDIES/COMMITMENTS	Discussion of any proposed remedies to address competition concerns or commitments to address serious doubts as to the compatibility of the notified operation.
8	CONCLUSION	The final decision and legal reasoning.
9	ANNEXES	Supporting documents, economic analysis, and third-party input.

Note. Adapted from European Commission merger decisions (2024)

Result Discussion

In this experiment, all three models were limited to 5000 features (TF-IDF tokens). Parameter tuning tested all possible parameter combinations five times using random data splits for training and testing. This study demonstrates that a text-based approach with NLP and ML can be used to predict EC merger decision outcomes using extracted text from official merger decisions.

The hybrid model was built using a stacking ensemble of Random Forest, SVM, and Naive Bayes classifiers. By training a meta-classifier on the predicted probabilities of these base models, the hybrid model achieved an overall accuracy of 98% when 3 base models were trained on the datasets containing 2.2K and 1.9K, outperforming all individual classifiers. This validates the benefit of combining models for complex regulatory decision modeling.

Principal Component analysis was also applied on all three base models to reduce dimensionality and improve performance while preserving as much variance as possible. Provided insights into better understanding of factors driving outcomes.

During the analysis, the main factors that typically lead to M&A rejections by the EC were also identified: **Reduction in Competition, Horizontal Overlaps, Vertical Foreclosure, Impact on Consumers.**

After training the model, performance was evaluated on the test set for base model as well as meta model. The following metrics were used to measure the model’s effectiveness.

Accuracy : The overall accuracy of the model in predicting the correct decision outcome (Approved, Conditionally Approved or Blocked).

Classification Report : The classification report provided detailed insights into Precision, Recall and F1-Score. This report provides statistics on the accuracy of classification models.

S. No.	Dataset	Model	Accuracy (%)	precision	recall	f1-score
1	additional_all_cases[2221]	Hybrid Model	98%	91%	99%	95%
		Random Forest	97%	90%	92%	91%
		SVM	93%	73%	82%	77%
		Naive Bayes	94%	72%	90%	80%

Figure 4
Performance Metrics of the Hybrid Machine Learning Model

Note. The hybrid model achieved high recall and accuracy across three outcome categories—approved, conditionally approved, and blocked—outperforming individual classifiers. Evaluation based on cross-validation results.

Confusion Matrix : The confusion matrix showed how well the model distinguished between the three decision outcomes. Confusion matrix (TP, TN, FP, FN) also checks the accuracy of classification models.

The confusion matrix in Table 3 summarizes the tenfold cross-validation results for 667 cases equally split between classes. Out of 96 cases ‘approved conditionally/unconditionally’, 87 were identified correctly and 9 were identified incorrectly. Out of 571 ‘Referral/Blocked, 570 cases were classified correctly, and 1 case was classified incorrectly. This suggests that the model performed better at predicting cases ‘Referral/Blocked’ than cases ‘approved conditionally/unconditionally’.

Table 3
Confusion Matrix from Ten-Fold Cross Validation

Actual Class	Predicted: Approved (TP/TN)	Predicted: Referral/Ban (FP/FN)
--------------	--------------------------------	------------------------------------

Approved Conditionally/Unconditionally	87 (True Negative)	9 (False Positive)
Referral/Ban	1 (False Negative)	570 (True Positive)

Note. Confusion matrix results based on 667 test cases using ten-fold cross-validation. TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative.

Hybrid model was found to be the best model among the other model for predicting mergers that were approved with or without conditions or referred/blocked.

This Hybrid model yielded a high recall of 98% to identify cases as highlighted in the Figure- 4 with serious anticompetitive effects and a relatively lower precision of 94% to avoid imposing unnecessary conditions on cases that do not have anticompetitive effects. It also used keywords on relevant markets, indicating its ability to capture classification nuances. Practical use requires a cost-benefit analysis to determine the optimal trade-off between recall and precision based on EC's specific objectives and priorities.

Conclusion and Future Enhancement

This study developed and evaluated a hybrid machine learning model to predict European Commission (EC) merger control decisions. By leveraging the strengths of three traditional classifiers Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB)—the proposed stacking ensemble approach significantly improved prediction accuracy, precision, recall, and F1-score. The hybrid model outperformed all individual classifiers, demonstrating the benefit of combining diverse predictive algorithms for complex regulatory decision-making.

Our results suggest that machine learning, particularly ensemble learning, can serve as an effective decision-support tool for legal and economic analysts involved in merger control cases. The hybrid model provides more consistent and interpretable outcomes that align with EC regulatory decisions, thereby enhancing transparency and predictive reliability in competition policy enforcement.

Additionally, the research reinforces the potential of AI to assist in antitrust evaluations, especially in identifying patterns within high-dimensional datasets involving market structure, industry codes, and merger characteristics.

Stacking with Neural networks based regulatory model could be explored to further improve performance and scalability. However, getting such a huge amount of data for such regulatory driven confidential data would be a challenging task.

Expanding the dataset to incorporate merger cases from other jurisdictions (e.g., U.S. FTC or Chinese MOFCOM) would enhance model generalizability and cross-border applicability.

References:

1. Competition: Merger control procedures.(n.d.). European commission. Retrieved April 13, 2024, from https://competition-policy.ec.europa.eu/mergers/procedures_en.
2. European Commission. (2023). *EU merger control overview*. Retrieved from <https://competition-policy.ec.europa.eu>
3. Competition Policy : Statistics on Mergers cases. (2025,February 28). European commission. https://competition-policy.ec.europa.eu/mergers/statistics_en
4. European Commission. (2004). Guidelines on the assessment of horizontal mergers Retrieved January 31, 2024, from <https://eur-lex.europa.eu/EN/legal-content/summary/guidelines-on-the-assessment-of-horizontal-mergers.html>
5. European Commission. (2004). Competition Policy Cases. Aaccessed on January 31, 2024, <https://competition-cases.ec.europa.eu/search?caseInstrument=M>

6. Jiang, T. (2021). Using Machine Learning to Analyze Merger Activity. *Frontiers in Applied Mathematics and Statistics*, 7, 649501. <https://doi.org/10.3389/fams.2021.649501>
7. Routledge, B. R., Sacchetto, S., & Smith, N. A. (2017). Predicting Merger Targets and Acquirers from Text. Carnegie Mellon University, Working Paper.
8. Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*. <https://doi.org/10.1086/688176>
9. Medvedeva, M., Vols, M., & Wieling, M. (2018). Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies in Europe (CELSE 2018)*.
10. Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28, 237 - 266. <https://doi.org/10.1007/s10506-019-09255-y>
11. Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management*, 56(5), 1618–1632. <https://doi.org/10.1016/j.ipm.2019.05.003>
12. Katz, D. M., Bommarito, M. J., & Blackman, J. (2020). A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 15(4), e0231765. <https://doi.org/10.1371/journal.pone.0231765>
13. Ravindran, R., Sharma, G., & Jain, R. (2021). Ensemble learning approaches for regulatory compliance and prediction in high-stakes domains. *Expert Systems with Applications*, 168, 114395. <https://doi.org/10.1016/j.eswa.2020.114395>
14. Arbo, A. (2023). Predicting Merger Decision Outcomes of the European Commission: A Natural Language Processing and Machine Learning Approach. Retrieved from <http://adellearbo.de/2023-06-08-predicting-merger-decision-outcomes/>
15. Mats A. Bergman, Maria Jakobsson, Carlos Razo. (2005). An econometric analysis of the European Commission's merger decisions. *International Journal of Industrial Organization*. <https://doi.org/10.1016/j.ijindorg.2005.08.006>.
16. Morzenti, G. (2023). Antitrust Policy and Innovation. University of Bocconi Department of Economics, from working paper https://www.dropbox.com/s/pxwr4wt9hh5rate/Antitrust_and_Innovation___Paper_V2.pdf?e=1&dl=0
17. European Commission. (2004). Competition Policy Cases. Aaccessed on January 31, 2024, <https://competition-cases.ec.europa.eu/search?caseInstrument=M>
18. Branting, K., Balhana, C., Pfeifer, C., Aberdeen, J., & Brown, B. (2020). Judges are from Mars, pro se litigants are from Venus: Predicting decisions from lay text. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9–11, 2020* (Vol. 334, p. 215). IOS Press.
19. European Commission. (2004). Competition Policy Cases. Retrieved January 31, 2024, from https://ec.europa.eu/competition/mergers/cases1/202307/M_10786_8967556_693_3.pdf