ISSN: 2282-4189

Research Center for European
Analysis and Policy

Jean Monnet Centre of Excellence on
EU Inclusive Open Strategic Autonomy

LUISS

Predicting AI patenting of European firms:
A Machine Learning approach

Francesco Bloise, Cristiana Fiorelli, Valentina Meliciani

Working Paper 4/2025

June 25, 2025

Project n. 101127624

Co-funded by
the European Union

"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Education and Culture Executive Agency. Neither the European Union nor the granting authority can be held responsible for them."

Predicting AI patenting of European firms: A Machine Learning approach

Francesco Bloise
Sapienza University of Rome, Italy

Cristiana Fiorelli
Sapienza University of Rome, Italy

Valentina Meliciani
LUISS University, Italy

Abstract

This paper investigates the firm-level predictors of artificial intelligence (AI) innovation in Europe using a machine learning (ML) approach. Drawing on a matched dataset that combines patent information from OECD REGPAT with firm-level financial and structural indicators from ORBIS Intellectual Property, we model the probability that a firm filed at least one AI-related patent at the European Patent Office (EPO) in 2020. AI patenting is identified using the World Intellectual Property Organization (WIPO) taxonomy. At the same time, the predictor set includes sectoral dummies, innovation and technological capabilities, balance sheet indicators, and interaction terms, resulting in a high-dimensional feature space. We compare traditional econometric models (Probit) with ML classifiers, including LASSO, Elastic Net, and Random Forest. Results reveal that ML algorithms significantly outperform Probit in identifying AI innovators, with Random Forest achieving the highest sensitivity (0.815) and balanced accuracy (0.854). These improvements are particularly valuable given the small share of firms performing AI patenting in the sample (3.7%). Variable importance analysis highlights ICT specialisation, patent stock, firm size, and market concentration as key drivers of AI innovation. Moreover, several interaction terms—such as those involving size, sector, and innovation intensity—emerge as critical in improving classification performance. These findings suggest that heterogeneities and nonlinear factors should be considered when predicting AI patenting. From a policy perspective, our results highlight that ML could help improve the tailoring of innovation policies by identifying firms with transformative potential. Supporting digital infrastructure, facilitating the diffusion of innovation beyond ICT-intensive hubs and tailoring strategies to firm size could make Europe's AI transformation more inclusive. Overall, this study demonstrates the analytical power of ML in uncovering complex innovation dynamics and providing actionable insights for scholars and policymakers.

Keywords: AI; innovation; firms; Machine Learning

JEL classification: O31; O34; C53

1. Introduction

Innovation is a key factor for firms, as it is the driver of long-term growth and international competitiveness (Dosi et al., 2015). It maximises a firm's ability to compete in global markets and survive the emergence of new competitors (Cefis and Marsili, 2006). Recently, technologies based on Artificial Intelligence (AI) have become increasingly prevalent. In this context, AI innovation has emerged as a strategic factor in enhancing firm productivity and promoting firm competitiveness and development (Damioli et al., 2021; da Silva Marioni et al., 2024). Accordingly, AI innovation has been positioned at the heart of strategic discussions on the industrial policies of European firms (European Commission, 2020).

Recent policy initiatives have emphasised the need for a coordinated European strategy on AI, calling for substantial public and private sector investment, stronger EU-wide governance, and a fully integrated digital single market. In this regard, Draghi (2024) outlines a possible strategy to close the innovation gap in a recent report on European competitiveness. It suggests that Europe vertically integrate AI technologies in key industrial sectors such as automotive, robotics, and pharmaceuticals. According to the report, targeted technological spending and cross-border collaborations are key factors in closing the innovation gap and improving Europe's AI capability. Therefore, to provide insights for designing effective support mechanisms and investment strategies across European industries, it is necessary to identify specific firms' characteristics that influence businesses' ability and willingness to achieve AI innovation. Although there has been a rapid growth in scholarly interest in AI, only a few contributions have explored which types of firms are most likely to engage in AI-related innovation (Martinelli et al., 2021; Igna and Venturini, 2023). Understanding the factors that enable—or hinder—AI innovation, such as firm size, sector, accumulation of digital human capital, prior experience with 4IR technologies, access to collaboration networks, and public support is crucial. First, it helps identify the barriers that prevent SMEs and "laggard" firms from adopting AI, thereby limiting the diffusion of productivity gains across the broader economy. Second, it provides actionable insights for policymakers on which levers to activate to foster a more inclusive and widespread adoption of AI technologies.

In contrast, most recent studies have examined the economic impact of AI in terms of productivity (Czarnitzki et al., 2023; da Silva Marioni et al., 2024), the labour market (Acemoglu and Restrepo, 2020; Damioli et al., 2023), and wage and income inequalities (Acemoglu, 2025). Previous evidence shows that a few large firms are responsible for a high proportion of patenting activity, primarily in the Information and Communication Technology (ICT) sector and clustered in a few tech hubs.

Regarding prediction tasks, growing literature explores machine learning (ML) to improve policy targeting and prediction. ML tools have been applied across diverse domains—including fiscal policy, poverty alleviation,

financial aid, credit access, and energy efficiency—offering more efficient and often more effective alternatives to traditional targeting methods. Studies have shown that ML can better identify beneficiaries in tax rebate programs (Andini et al., 2018), optimise residential energy retrofits (Christensen et al., 2024), and enhance the allocation of public credit guarantees (Andini et al., 2022). In development settings, ML has been used with mobile phone data to improve poverty targeting in Afghanistan (Aiken et al., 2023) and to assess the feasibility of predicting entrepreneurial success (McKenzie and Sansone, 2019). Work by Athey et al. (2025) illustrates the value of combining predictive and causal ML for targeting educational behavioural interventions. These contributions highlight the potential of ML to improve policy design and delivery, though issues of interpretability, fairness, and transparency remain central. Ludwig and Mullainathan (2024) emphasise the potential of ML for targeting and hypothesis generation. By uncovering patterns humans might overlook, ML can inspire new, interpretable research questions from complex data. These characteristics expand its role beyond prediction into the scientific discovery process. The field is expanding rapidly, with evidence supporting both the promise and limitations of ML as a tool for public policy.

In this paper, we use ML to predict AI innovation at the firm level in a sample of European countries. To this end, we construct our dataset using ORBIS Intellectual Property (ORBIS-IP), a matched firm–patent database. We focus on active financial and non-financial firms headquartered in the EU-15 that filed at least one EPO patent in 2020. AI-innovating firms are identified by linking ORBIS-IP with the OECD REGPAT database, which provides granular data on global patent applications.

We perform a classification analysis using state-of-the-art ML algorithms (i.e., LASSO, Elastic net, Random Forest, Gradient Boosting Machine). We exploit a high-dimensional vector of potential predictors, their lagged values, and high-order nonlinearities among predictors to maximise the capability of our ML models to predict our target variable accurately. Unlike previous evidence based on standard regression models, which are forced to decide ex-ante a limited number of predictors, our approach can capture relevant predictors without assuming a linear functional form and limiting our analysis to predictors observed at a given time. Therefore, our data-driven approach helps identify relevant predictors of AI innovation that have not been analysed in previous studies from a theoretical or empirical point of view and by capturing potential nonlinearities among predictors.

Our results show that ML models—particularly Random Forest—outperform traditional econometric approaches in predicting AI innovation, achieving high sensitivity and balanced accuracy despite the rarity of AI patenting in the sample. Relevant predictors of AI patenting are ICT specialisation, prior patent activity,

firm size, technological specialisation, and several relevant interaction effects. These findings underscore the complexity and heterogeneity of AI adoption processes and the value of data-driven approaches.

The remainder of the paper is organised as follows: Section 2 reviews the relevant literature; Section 3 presents the data and methodological framework; Section 4 provides descriptive evidence; Section 5 discusses the empirical results; and Section 6 concludes the paper with policy implications and directions for future research.

2. Literature review

The literature has mainly focused on AI's impact on productivity, labour demand, income distribution, and the heterogeneous diffusion of technology, given the structural differences in capabilities, infrastructure, and institutions across sectors and regions (Guarascio et al., 2025).

Initial studies examined the effects of changes in the capital-to-labour ratio caused by automation and robotics on the labour market. These studies demonstrate that mid-skilled workers performing manual, highly routinised tasks are particularly vulnerable to displacement by industrial robots (Autor et al., 2003; Acemoglu & Restrepo, 2020). This phenomenon is generally characterised as the labour-saving effect of automation. On the other hand, Damioli et al. (2023) investigated the labour-friendly nature of AI technologies, supporting the hypothesis that the employment impact of AI is larger and more significant than the job creation effect of other innovative activities. Regarding 'AI exposure' (Felten et al., 2018), some studies have found that employment shares tend to increase in AI-exposed occupations in Europe, particularly those characterised by a relatively high proportion of younger and skilled workers (Albanesi et al., 2023). However, the effect varies geographically (Guarascio and Reljic, 2024).

Although estimates of the contribution of AI to aggregate productivity remain uncertain (Acemoglu, 2025), a growing body of patent-based evidence documents a strong link between AI innovation and productivity at the firm level. In the United States, for example, Alderucci et al. (2020) demonstrate that companies that patent AI technologies are more productive than their non-AI counterparts. Yang (2022) analysed 600 Taiwanese electronics firms from 2002 to 2018, finding that a 10 per cent rise in AI patents boosts overall productivity by about 5 per cent. Using a panel of worldwide firms from 2000 to 2016, Damioli et al. (2021) found that patenting in AI was associated with a 3 per cent increase in labour productivity, with the gains particularly pronounced among SMEs and service sector companies. Furthermore, Benassi et al. (2022) demonstrate that accumulating Fourth Industrial Revolution (4IR) knowledge increases output per worker and multifactor productivity. The largest effects stem from AI, wireless technologies, cognitive computing, and big data analytics, especially for

firms with prior experience in pre-4IR domains or those that adopted these technologies early. Recently, da Silva Marioni et al. (2024) used a difference-in-differences quasi-experimental approach to estimate the causal effect of AI innovation on the productivity of 15 European countries between 2011 and 2019. They found that company productivity increased by an average of 6–11%, with the greatest gains made by technologically backward firms.

Given the evidence above showing that adopting AI delivers significant productivity gains, it is important to identify the characteristics of firms more likely to innovate in this field. As highlighted by the existing literature, most AI patents are held by a small group of large, established companies, mainly ICT firms clustered in just a few global tech hubs. Dernis et al. (2019) highlight that 75% of AI patents worldwide are filed by top R&D performers, with software companies and IT service providers now surpassing traditional high-tech manufacturers in AI-related innovations. Furthermore, Klinger et al. (2020) find that AI-related scientific publications are dominated by large high-tech firms exhibiting decreasing diversification in their product portfolios. This pattern suggests that AI development is becoming more concentrated, potentially limiting the diffusion of AI capabilities across industries and restricting the ability of smaller firms to enter the market (Fanti et al., 2022). More recently, Igna and Venturini (2023) show that the probability of inventing AI is systematically higher for major innovators already active in fields such as ICT.

Regarding patent productivity, AI innovation presents strong dynamic returns from the knowledge earlier developed in network and communication technologies. Cumulative knowledge dynamics also emerge in territorial analyses. For example, Xiao and Boschma (2023) examined AI innovation capacity across 233 European regions from 1994 to 2017. They found that regions with the highest shares of AI patents already possessed a strong ICT knowledge base, providing evidence of path-dependent diffusion. These findings are consistent with those of Buarque et al. (2020), who demonstrated that regions most successful in AI were precisely those in which AI-related technologies were deeply embedded within the local knowledge space. Therefore, knowledge concentration concerns not only the industrial sector but also the specific local characteristics of the area where a firm is based.

Differently from the existing literature (Igna and Venturini, 2023) that relies on parametric models and focuses on the role of past ICT capabilities and firm-level knowledge accumulation as key drivers of AI patent productivity, our approach adopts a high-dimensional, non-parametric ML framework that allows us to uncover complex, nonlinear interactions among a much broader set of predictors, without imposing strong a priori assumptions on the functional form of the relationships. Unlike traditional econometric techniques such as Probit models—which require a predefined and limited set of covariates and assume linearity and additivity

in the effects—ML algorithms can flexibly capture intricate dependencies and interactions across hundreds of features, including higher-order and interaction terms. This flexibility improves predictive performance, especially in contexts such as AI innovation, where relationships are expected to be highly heterogeneous and nonlinear. Moreover, ML tools allow us to identify novel, previously overlooked predictors and interaction effects that would be difficult to detect through conventional estimation approaches.

3. Data and Methodology

3.1 Data

The dataset is constructed starting from ORBIS Intellectual Property (ORBIS-IP), Bureau van Dijk's new matched patent–firm database that links granular patent records with firm-level information for around 110 million companies worldwide (Benassi et al., 2022). From this dataset, following Igna and Venturini (2023), we selected all the financial and non-financial active corporations that filed at least one patent application with the European Patent Office (EPO) in 2020 (priority year).[1] All selected firms are headquartered in the EU-15 countries (i.e., Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden and the United Kingdom). We remove all observations whose ORBIS consolidation flag indicates that reliable financial statements are unavailable. Concretely, we drop firms with consolidation codes equal to "NF" (no financials) and "LF" (limited financials, i.e. only sales and total assets are reported), as well as those for which the code is missing. These entities lack the balance-sheet variables required for our analysis; retaining them would introduce measurement error without adding usable information. Moreover, we remove firms for which the necessary balance sheet information for our analysis is missing to obtain a final sample of 3,243 firms.

We classify AI-innovating and non-AI-innovating firms by linking ORBIS-IP to the microdata from the OECD REGPAT database (January 2024 edition). The latter, which contains information on individual patent applications worldwide, collects data from the European Patent Office (EPO) and the Patent Cooperation Treaty (PCT). Matching is performed on the publication numbers of applications filed at the EPO by European applicants. The literature recognises two principal approaches for identifying AI patents. The first uses the EPO taxonomy (EPO 2017, Annex 1), which aggregates Cooperative Patent Classification (CPC) codes spanning the broader 4IR technology space (Benassi et al., 2022; Igna and Venturini, 2023). This scheme may over-count AI patents because 4IR encompasses many fields beyond AI. The second is the PATENTSCOPE Artificial

[1] We selected 2020 for the analysis because patent data is generally recorded with three to four years of delay, and our most recent database update is in 2024.

Intelligence Index developed by the World Intellectual Property Organization (WIPO, 2019), which provides key phrases, IPC (International Patent Classification), and CPC codes used for capturing core AI technologies (Damioli et al., 2021; da Silva Marioni et al., 2024; Kopka and Fornahl, 2024). The index is divided into two segments. In our paper, we follow the approach of Xiao and Boschma (2023) and use only the CPC codes from the first segment of the WIPO classification.[2] Starting from this information, we construct a dummy variable *AI* that equals one if the firm filed at least one patent application containing an AI-related CPC code in 2020 and 0 otherwise.

We employ a broad set of predictors related to sectoral affiliation, accumulated technological capabilities, structural characteristics, and balance sheet indicators to capture the main factors influencing a firm's ability to innovate in AI.

A detailed description of predictors is provided in Table 1. The first group of predictors concerns the firms' innovative and technological capabilities. We include the yearly number of patents filed by each firm (*m_nrpat*), regardless of whether they are related to AI, to capture their general propensity to innovate. Building on this, we also consider the degree of technological specialisation using the Herfindahl-Hirschman Index (*m_HHI*). This index, which ranges from 0 to 1 (with 1 indicating maximum concentration), is calculated based on the distribution of patents across the main CPC macro-categories (A to Y), representing different technological domains.

Furthermore, we include the dummy variable *ICT* that takes the value 1 when the firm filed at least one patent application in the fields of ICT between 2016 and 2019. To identify patent applications related to ICT, we rely on the J-tag classification system proposed by Inaba and Squicciarini (2017). As this taxonomy includes categories overlapping with AI codes, we exclude these codes from the broader set of ICT-related codes (Igna and Venturini, 2023).

Recent work highlights the growing complementarity between digital and green technologies—the so-called twin transition (Diodato et al., 2023; Montresor and Vezzani, 2023). New digital technologies can further enhance the effectiveness of green technologies, thereby increasing their overall environmental benefits (see Biggi et al., 2025). Companies already working on green solutions are likely to adopt AI as well. To embed this perspective, we add a predictor that can capture the propensity of green innovation. Specifically, we create the dummy *green* that takes the value one if the firm filed at least one patent in the green area between 2016 and

[2] We verify the strength of our results by employing the EPO 4IR taxonomy as a robustness check.

2019. Following Fabrizi et al. (2025), we apply the "Y02/Y04S tagging scheme" developed by the EPO to identify the green applications.

The second set of predictors describes each firm's structural characteristics. We also add the industry affiliation by introducing a set of dummies (*nace_*) that assign every firm to one of the different NACE Rev. 2 sections (agriculture, manufacturing, construction, information and communication, and so on).[3] Next, we introduce a dummy variable *group* to identify whether a company belongs to a wider corporate group. Group affiliation can provide access to shared R&D facilities, internal capital markets, and cross-firm knowledge flows, thereby increasing innovative capability. Finally, we proxy firm size with four employment-class dummies that follow the Eurostat definition: *empl_1* flags micro-enterprises with up to 9 employees; *empl_2* covers small firms with 10–49 employees; *empl_3* captures medium-sized firms with 50–249 employees; and *empl_4* identifies large enterprises with 250 or more employees.

The last set of predictors accounts for the balance sheet indicators. First, we consider the size of a firm's balance sheet with total assets (*m_tot_asset*) and capture the flow dimension of its activity with operating revenues or turnover (*m_op_rev*). Intangible assets (*m_intangible*) are a stock measure for past innovation spending and accumulated knowledge capital. At the same time, the profitability enters through the EBITDA margin (*m_EBITDA*), signalling the cash a firm can generate internally to finance innovative projects. We also include measures relating to financial position, such as the stock of long-term debt (*m_Longdebt*) and the current ratio (*m_currentr*), which compares current assets to current liabilities and indicates a firm's ability to cover its short-term obligations. Finally, we account for market power measured by a proxy of the Lerner index (*lerner_index*), defined as operating profit over operating revenue, which captures the extent to which firms can set price above cost (Aghion et al., 2005).

[3] Table A1 of the Appendix provides detailed information about specific NACE categories included in the vector of predictors.

Table 1: Variable description and sources.

Variable	Description	Source
AI	Dummy: 1 if the firm filed at least one patent application containing an AI-related CPC code in 2020; 0 otherwise	OECD REGPAT database; own elaboration
ICT	Dummy: 1 if the firm filed at least one patent application ICT between 2016 and 2019; 0 otherwise	OECD REGPAT database; own elaboration
m_HHI	Herfindahl-Hirschman Index on technological specialisation; average between 2016 and 2019	OECD REGPAT database; own elaboration
m_nrpat	Number of patents filed by each firm; average between 2016 and 2019	ORBIS-IP
green	Dummy: 1 if the firm filed at least one patent application in the green area between 2016 and 2019; 0 otherwise	OECD REGPAT database; own elaboration
nace_	Industry affiliation, NACE Rev. 2 sections; dummies	ORBIS-IP database; own elaboration
group	Belonging to a wider corporate group, dummy	ORBIS-IP database; own elaboration
empl_	Firm size with four employment classes; dummies	ORBIS-IP database; own elaboration
m_tot_asset	Total assets; average between 2016 and 2019	ORBIS-IP database
m_op_rev	Operating revenues (turnover); average between 2016 and 2019	ORBIS-IP database
m_intangible	Intangible assets; average between 2016 and 2019	ORBIS-IP database
m_EBITDA	EBITDA margin; average between 2016 and 2019	ORBIS-IP database
m_Longdebt	Long-term debt; stock; average between 2016 and 2019	ORBIS-IP database
m_currentr	Current ratio; average between 2016 and 2019	ORBIS-IP database
lerner_index	Operating profit over operating revenue; average between 2016 and 2019	ORBIS-IP database; own elaboration

3.2 Methodology

To predict firm-level adoption of artificial intelligence (AI) technologies, we model the probability that a given firm files at least one AI-related patent in 2020. The dependent variable is defined as:

$$AI_{f,2020} = f(X_{f,2020-1}) + \varepsilon_{f,2020,} \quad (1)$$

where $AI_{f,2020}$ is a binary indicator equal to 1 if firm x files an AI patent in 2020 and 0 otherwise. The function f(.) denotes an unknown relationship between AI patenting and firm-level characteristics observed between 2016 and 2019, and ε_x is an idiosyncratic error term.

The dependent variable's binary nature justifies comparing different classification algorithms. We estimate a baseline Probit model, followed by machine learning (ML) methods, including LASSO, Elastic Net, and

Random Forest. LASSO and Elastic Net are regularisation-based estimators allowing variable selection and shrinkage, while Random Forest is a non-parametric ensemble learning technique.

To improve flexibility and capture nonlinearities, we expand the feature space to include squared terms and all possible two-way interactions for LASSO and Elastic Net, resulting in 423 candidate predictors. Random Forest automatically incorporates high-order nonlinearities among predictors. Mullainathan and Spiess (2017) suggested to adopt a three-step approach. First, we randomly divide our sample into a training sample to calibrate and estimate our algorithms and a test sample. Second, we exploit the training sample to calibrate our ML algorithms using 5-fold cross-validation. Third, we estimate our calibrated model to obtain our-of-sample scores in the test sample and identify relevant predictors of AI innovation.

We rely on two key performance metrics, sensitivity (recall) and balanced accuracy, to assess prediction quality. These are computed as follows:

$$Sensitivity = TP / (TP + FN) \quad (2)$$

$$Balanced\ accuracy = (Sensitivity + Specificity) / 2, \quad (3)$$

where TP is the number of true positives, FN is the number of false negatives, and specificity is the true negative rate. To mitigate overfitting and underfitting, we apply 5-fold cross-validation.[4] Each model is calibrated on a training set (80% of the data) and evaluated on a hold-out set (20%) to obtain unbiased performance estimates.

4. Descriptive Evidence

The sample consists of 3,243 firms in EU-15 countries that filed at least one patent at the EPO in 2020. Only 3.7% are identified as AI innovators. This low incidence underscores the relevance of using high-sensitivity classification models to identify rare but significant cases of technological leadership.

The average firm in the dataset exhibits high heterogeneity in size, sector, and technological capabilities. The mean ICT specialisation score is 0.101, with a standard deviation of 0.302, suggesting that only a small subset

[4] LASSO and Elastic Net are calibrated by testing 50 different regularisation parameter (lambda) values. For Elastic Net, the tuning also included three different values of the mixing parameter (alpha). Random Forest is calibrated after a variable pre-selection step based on the non-zero coefficients identified by LASSO. The calibration considers different values for the proportion of features used at each split and the number of splits, with the number of trees fixed at 500.

of firms maintains a strong ICT orientation. Similarly, the average firm has 120 patents (mean of m_nrpat), but this indicator has high dispersion (standard deviation equals 643), implying a skewed distribution. Notably, 32% of firms are active in green innovation, and 23% belong to corporate groups, signalling heterogeneity in firm structure and innovation strategy. The sectoral distribution shows that over 66% of firms are in the manufacturing sector (nace_3), reflecting the dominant industrial cluster in the sample. The next most prevalent sectors are professional, scientific and technical activities (nace_12), followed by transport and storage (nace_8). Finally, firm size, proxied by employment class, shows that nearly half (48.9%) of the firms belong to the largest class (empl_4), while only 9.1% are micro-enterprises (empl_1). These results confirm that larger firms are overrepresented among patent filers, which aligns with prior literature on the innovation-size nexus.

Table 2: Descriptive evidence

Variable	**Mean**	**Std. Dev.**
AI	0.037	0.188
ICT	0.101	0.302
green	0.320	0.467
m_nrpat	120.203	643.330
nace_1	0.002	0.046
nace_2	0.004	0.061
nace_3	0.660	0.474
nace_4	0.007	0.086
nace_5	0.003	0.053
nace_6	0.003	0.051
nace_7	0.063	0.243
nace_8	0.063	0.242
nace_9	0.033	0.179
nace_10	0.003	0.053
nace_11	0.002	0.046
nace_12	0.163	0.370
nace_13	0.015	0.122
nace_14	0.000	0.018
nace_15	0.003	0.053
nace_16	0.009	0.096
nace_17	0.000	0.018
nace_18	0.003	0.053
empl_1	0.091	0.288
empl_2	0.148	0.355
empl_3	0.272	0.445
empl_4	0.489	0.500
Obs.	3,243	3,243

Notes: Authors' elaborations.

5. Results

This section presents the empirical findings from the estimation of traditional econometric and ML models (Table 3). Probit achieves a balanced accuracy of 0.523 and a sensitivity of just 0.083, indicating a minimal ability to identify actual AI innovators. These results highlight the challenge of using traditional models in the context of rare events and high-dimensional predictors.

In contrast, the LASSO and Elastic Net models significantly outperform the Probit specification. Both yield a sensitivity of 0.750 and a balanced accuracy of 0.812. This improvement can be attributed to their ability to penalise irrelevant predictors and select relevant predictors, including nonlinear and interaction effects. However, the best-performing algorithm is Random Forest, which achieves a sensitivity of 0.815 and a balanced accuracy of 0.854. These gains highlight the superiority of nonparametric models in capturing complex, nonlinear relationships between firm characteristics and AI adoption.

Table 3: predictive performance across models

Model	**Sensitivity**	**Balanced Accuracy**
Probit	0.083	0.523
LASSO	0.750	0.812
Elastic Net	0.750	0.812
Random Forest	0.815	0.854

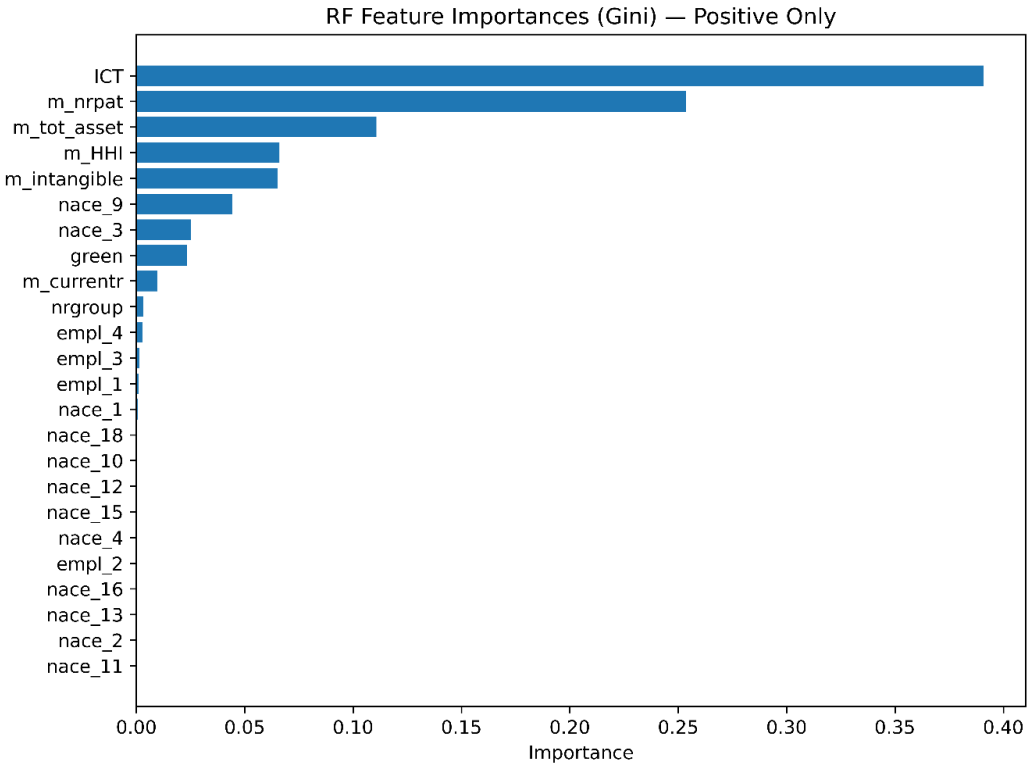
Notes: Authors' elaborations.

Beyond prediction, we examine variable importance using the Random Forest model. Importance is assessed using the Gini impurity reduction metric and permutation-based performance loss.[5] Both approaches converge on a consistent set of top predictors: ICT specialisation, patent stock (m_nrpat), firm size (empl_4), and technological specialisation (m_HHI). The stock of prior patents and a firm's intangible capital gives rise to strong dynamic increasing returns (learning-by-doing) and powerful knowledge complementarities. Companies with codified intellectual property and tacit assets (software routines, brands, organisational know-how) enjoy a cost advantage in subsequent AI invention because they can recombine these capabilities at a lower marginal cost. As underlined by Igna and Venturini (2023), network effects amplify these cumulative

[5]Feature importance in Random Forest is assessed using two methods. The first is the Mean Decrease in Impurity (Gini importance), which measures how much each variable contributes to reducing node impurity across all trees in the forest. The second is permutation importance, which evaluates the change in model performance when the values of a single feature are randomly permuted—thus breaking its relationship with the outcome—while keeping all other features unchanged. Permutation importance has been computed over 50 random replications to ensure robustness and account for variability due to random shuffling.

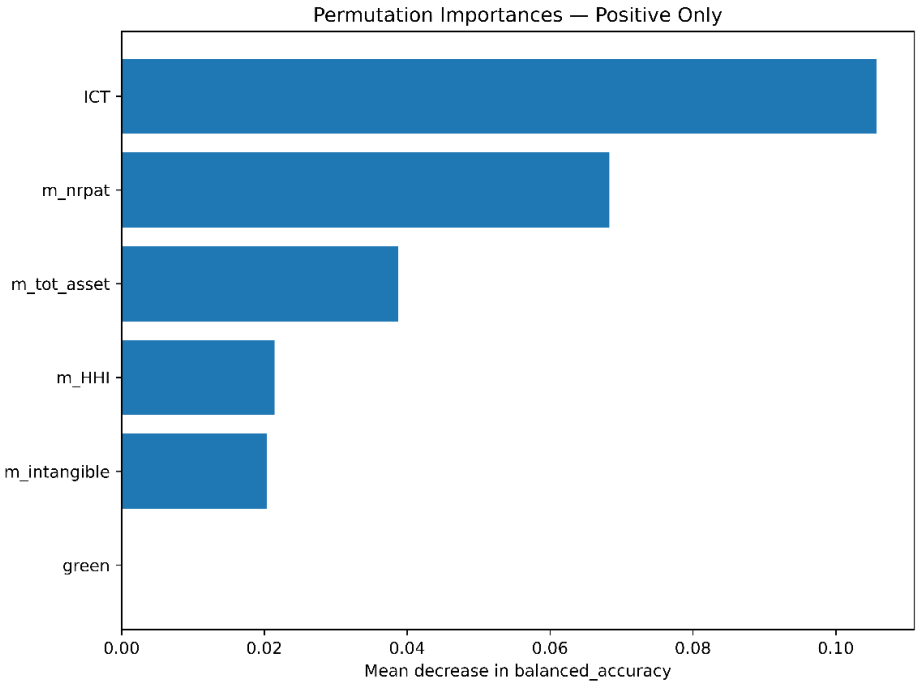
mechanisms within ICT, boosting the probability of AI patenting and reinforcing path dependence. Technological specialisation matters as well.

Figure 1: Feature importance from Random Forest



Notes: Authors' elaborations

Figure 2: Permutation importance from Random Forest



Notes: Authors' elaborations

To enhance interpretability, we also extract standardised coefficients from the LASSO model. Only predictors with coefficients greater than 0.05 in absolute value are retained in Table 4. The leading variable, ICT, shows a coefficient of 0.376, confirming its strong association with AI patenting. Other influential features include interaction terms between employment class and sector (e.g., Xempl_4ICT = 0.200) and between knowledge accumulation and sector (e.g., Xm_intangiblenace_10 = 0.075). Regarding innovation capabilities in specific sectors, we find strong evidence of the influence of green patenting knowledge on AI innovation. A positive coefficient for green × nace_3 (0.126) indicates that manufacturing firms already innovating in environmental technologies are more likely to innovate in AI, underscoring the "twin transition" in which digital and green advances reinforce one another. Scale also matters: the large-firm dummy interacted not only with ICT innovation but also with green innovation, which is strongly positive.

Table 4: selected coefficients from LASSO

Variable	**Estimated coefficient**
ICT	0.376
Xempl_4ICT	0.200
Xm_HHInace_9	0.141
Xgreennace_3	0.126
nace_9	0.122
Xm_nrpatgreen	0.096
Xempl_4nace_16	0.087
Xm_intangiblegreen	0.083
Xempl_2nace_13	0.076
Xm_intangiblenace_10	0.075
XICTnace_12	0.070
Xm_intangiblenace_15	0.069
Xempl_1nace_1	0.067
Xempl_4green	0.065
Xnrgroupnace_16	0.065
m_nrpat	0.065
Xm_nrpatnace_4	0.062
Xempl_2nace_9	0.059
Xempl_4m_nrpat	0.052
sqrm_HHI	-0.074
m_HHI	-0.102

Notes: All variables with the prefix '*X*" are interaction terms. A variable whose name starts with '*sqr*' is the square of a predictor.

Moreover, m_HHI and its squared value exhibit negative coefficients, suggesting a nonlinear effect of technological specialisation on AI innovation propensity. This result implies that AI invention thrives on a diversified mix of heterogeneous knowledge and competencies (Igna and Venturini, 2023). However, the picture changes when specialisation interacts with a specific sector that accounts for the information and communication (i.e., Xm_HHInace_9= 0.141). Specialised firms operating within information and communication technologies enjoy a markedly higher probability of innovating in AI. By accommodating such nonlinearities and interaction effects, machine-learning estimators reveal a far richer mapping from firms' knowledge profiles to AI invention than traditional linear models could ever detect. At the opposite end of the size spectrum, empl_1 × nace_1 (0.067) reveals that even agricultural micro-enterprises have above-average odds of AI patenting. This result aligns with the recent surge in 'precision farming' and 'digital farming' solutions, such as UAV-based multispectral imaging for detecting crop stress, convolutional neural network

classifiers for recognising weeds, and reinforcement learning-driven variable-rate irrigation systems. These solutions are employed in agriculture to increase production efficiency.

Overall, these results demonstrate the added value of ML methods in uncovering complex patterns that would be difficult to detect using traditional techniques. They also confirm the relevance of firm-level heterogeneity, sectoral context, and historical innovation capacity in shaping AI adoption.

6. Conclusions and Policy Suggestions

This paper has demonstrated that ML algorithms can effectively predict AI innovation at the firm level using a rich set of balance sheets and sectoral and innovation-related indicators. Based on the presence of AI patent applications in 2020, the classification task posed significant challenges due to class imbalance and nonlinearity in firm characteristics. Traditional econometric models, such as Probit, failed to capture these complexities and yielded very low predictive power.

Conversely, ML techniques such as Random Forest and regularisation-based models, are substantially more accurate. Random Forest achieved the highest sensitivity and balanced accuracy, while LASSO provided insight into the magnitude and direction of key predictors. These findings confirm that ML methods are well suited to identifying rare yet economically significant innovation patterns across firms.

Importantly, this study highlights superior predictive performance and sheds light on the key drivers of AI innovation. ICT specialisation is the most consistent determinant, reflecting the need for digital infrastructure and absorptive capacity. Furthermore, the presence of significant interaction terms indicates that the impact of a predictor frequently hinges on the levels of the others.

From a policy perspective, these results offer important suggestions. Firstly, they support the use of ML in designing more targeted innovation policies. By identifying firms with a high probability of adopting AI, policymakers can allocate resources more effectively and provide targeted support in the form of R&D grants and tax credits. This approach ensures that public investments focus on firms with strong transformative potential.

Secondly, insights from LASSO coefficients enable policymakers to understand who innovates and why. For example, the significant involvement of ICT-intensive or green-intensive sectors highlights the importance of dynamic returns of knowledge. At the same time, the interaction between firm size and innovation indicators highlights the need for differentiated strategies for SMEs versus large enterprises.

Thirdly, the results suggest that AI innovation remains highly concentrated and path-dependent, which echoes previous findings on technological lock-ins. Supporting the diffusion of AI therefore requires addressing the systemic barriers – such as skill gaps, financing constraints and network externalities – that hinder late adopters.

Future research should build on this work in three directions. First, alternative patent taxonomies, such as the EPO 2017 classification, could be employed to validate the robustness of the model. Second, restricting the analysis to firms with no AI patenting history from 2016 to 2019 would help identify predictors of first-time adoption. Third, causal inference techniques, including Double Machine Learning (DML), could be used to estimate the impact of public interventions.

This paper advances the methodological and substantive understanding of AI innovation in Europe. It shows that ML methods are robust classifiers and valuable tools for revealing the empirical structure of innovation adoption. Using these methods could greatly improve researchers' and policymakers' ability to predict, support and expand the next wave of technological change.

References

Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy*, *40*(121), 13-58.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, *128*(6), 2188-2244.

Aghion, P., Bloom, N., Blundell, R., Griffith, R., & Howitt, P. (2005). Competition and innovation: An inverted-U relationship. *The quarterly journal of economics*, *120*(2), 701-728.

Aiken, E. L., Bedoya, G., Blumenstock, J. E., & Coville, A. (2023). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, *161*, 103016.

Albanesi, S., Da Silva, A. D., Jimeno, J. F., Lamo, A., & Wabitsch, A. (2023). *New technologies and jobs in Europe* (No. w31357). National Bureau of Economic Research.

Alderucci, D., Branstetter, L., Hovy, E., Runge, A., & Zolas, N. (2020, January). Quantifying the impact of AI on productivity and labor demand: Evidence from US census microdata. In *Allied Social Science Associations–ASSA 2020 annual meeting*.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization*, *156*, 86-102.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, *118*(4), 1279-1333.

Athey, S., Keleher, N., & Spiess, J. (2025). Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal. *Journal of Econometrics*, 105945.

Benassi, M., Grinza, E., Rentocchini, F., & Rondi, L. (2022). Patenting in 4IR technologies and firm performance. *Industrial and Corporate Change*, *31*(1), 112-136.

Biggi, G., Iori, M., Mazzei, J., & Mina, A. (2025). Green Intelligence: The AI content of green technologies. *Eurasian Business Review*, 1-38.

Buarque, B. S., Davies, R. B., Hynes, R. M., & Kogler, D. F. (2020). OK Computer: the creation and integration of AI in Europe. *Cambridge Journal of Regions, Economy and Society*, *13*(1), 175-192.

Cefis, E., & Marsili, O. (2006). Survivor: The role of innovation in firms' survival. *Research Policy*, *35*(5), 626-641.

Christensen, P., Francisco, P., Myers, E., Shao, H., & Souza, M. (2024). Energy efficiency can deliver for climate policy: Evidence from machine learning-based targeting. *Journal of Public Economics*, *234*, 105098.

Czarnitzki, D., Fernández, G. P., & Rammer, C. (2023). Artificial intelligence and firm-level productivity. *Journal of Economic Behavior & Organization*, *211*, 188-205.

da Silva Marioni, L., Rincon-Aznar, A., & Venturini, F. (2024). Productivity performance, distance to frontier and AI innovation: Firm-level evidence from Europe. *Journal of Economic Behavior & Organization*, *228*, 106762.

Damioli, G., Van Roy, V., & Vertesy, D. (2021). The impact of artificial intelligence on labor productivity. *Eurasian Business Review*, *11*, 1-25.

Damioli, G., Van Roy, V., Vertesy, D., & Vivarelli, M. (2023). AI technologies and employment: micro evidence from the supply side. *Applied Economics Letters*, *30*(6), 816-821.

Dernis, H., Gkotsis, P., Grassano, N., Nakazato, S., Squicciarini, M., van Beuzekom, B., & Vezzani, A. (2019). *World corporate top R&D investors: Shaping the future of technologies and of AI* (No. JRC117068). Joint Research Centre.

Diodato, D., Huergo, E., Moncada-Paternò-Castello, P., Rentocchini, F., & Timmermans, B. (2023). Introduction to the special issue on "the twin (digital and green) transition: handling the economic and social challenges". *Industry and Innovation*, *30*(7), 755-765.

Dosi, G., Grazzi, M., Moschella, D. (2015). Technology and costs in international competitiveness: From countries and sectors to firms. *Research Policy*, 44 (10), 1795-1814.

Draghi, M. (2024), The future of European competitiveness, Technical report, European Commission.

EPO (2017). Patents and the Fourth Industrial Revolution. Technical report, European Patent Office

European Commission (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust. European Union. Retrieved from https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Fabrizi, A., Fiorelli, C., & Meliciani, V. (2025). The role of green networks for environmental innovation in European regions. *Journal of Industrial and Business Economics*, 1-32.

Fanti, L., Guarascio, D., & Moggi, M. (2022). From Heron of Alexandria to Amazon's Alexa: a stylised history of AI and its impact on business models, organisation and work. *Journal of Industrial and Business Economics*, *49*(3), 409-440.

Felten, E. W., Raj, M., & Seamans, R. (2018, May). A method to link advances in artificial intelligence to occupational abilities. In *AEA Papers and proceedings* (Vol. 108, pp. 54-57). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Guarascio, D., & Reljic, J. (2025). AI and Employment in Europe. *Economics Letters*, *247*, 112183.

Guarascio, D., Reljic, J., & Stöllinger, R. (2025). Diverging paths: AI exposure and employment across European regions. *Structural Change and Economic Dynamics*, *73*, 11-24.

Klinger, J., Mateos-Garcia, J. C., & Stathoulopoulos, K. (2020). A Narrowing of AI Research?. *Available at SSRN 3698698*.

Kopka, A., & Fornahl, D. (2024). Artificial intelligence and firm growth—catch-up processes of SMEs through integrating AI into their knowledge bases. *Small Business Economics*, *62*(1), 63-85.

Igna, I., & Venturini, F. (2023). The determinants of AI innovation across European firms. Research Policy, 52(2), 104661.

Inaba, T., & Squicciarini, M. (2017). ICT: A new taxonomy based on the international patent classification. *OECD Science, Technology and Industry Working Papers*, *2017*(1), 1.

Ludwig, J., & Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, *139*(2), 751-827.

McKenzie, D., & Sansone, D. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. Journal of Development Economics, 141, 102369.

Montresor, S., & Vezzani, A. (2023). Digital technologies and eco-innovation. Evidence of the twin transition from Italian firms. *Industry and Innovation*, *30*(7), 766-800.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

WIPO (2019). WIPO Technology Trends 2019: Artificial Intelligence. Technical report.

Xiao, J., & Boschma, R. (2023). The emergence of artificial intelligence in European regions: the role of a local ICT base. *The Annals of Regional Science*, *71*(3), 747-773.

Yang, C. H. (2022). How artificial intelligence technology affects productivity and employment: firm-level evidence from Taiwan. *Research Policy*, *51*(6), 104536.

Appendix

Table A1: Description of the NACE Rev. 2 sections

Variable Name	Description
nace_1	Agriculture, forestry and fishing
nace_2	Mining and quarrying
nace_3	Manufacturing
nace_4	Electricity, gas, steam and air conditioning supply
nace_5	Water supply, sewerage, waste management
nace_6	Construction
nace_7	Wholesale and retail trade; repair of motor vehicles and motorcycles
nace_8	Transportation and storage
nace_9	Information and communication
nace_10	Financial and insurance activities
nace_11	Real estate activities
nace_12	Professional, scientific and technical activities
nace_13	Administrative and support service activities
nace_14	Public administration and defence; compulsory social security
nace_15	Education
nace_16	Human health and social work activities
nace_17	Arts, entertainment and recreation
nace_18	Other service activities