Data-Driven Policy: Forecasting the Socioeconomic Impact of Industrial Automation Using Machine Learning

LAWRENCE A. FARINOLA[1,2], JEAN-EUDES ASSOGBA[1,2], MAHOUGNON B. M. ASSOGBA[1,2]
[1]Department of Software Engineering
Faculty of Architecture and Engineering
Rauf Denktas University
Mersin 10 via
TÜRKIYE

[2]Center of Excellence for Interdisciplinary AI and Data Science Research
Rauf Denktas University
Mersin 10 via
TÜRKIYE

Abstract: - As industrial automation accelerates across diverse sectors, its socioeconomic repercussions remain complex and uncertain, particularly on employment, income distribution, and macroeconomic stability. This study proposes a data-driven framework that integrates labour-market microdata, industry performance metrics, and task-level automation-risk indices to forecast changes in key socioeconomic indicators. Five state-of-the-art machine learning algorithms—Random Forest, XGBoost, CatBoost, LightGBM, and TabNet—were trained and compared. These models capture complex, nonlinear interactions in socioeconomic data. Among them, Random Forest demonstrated the best predictive performance, achieving the lowest RMSE (1560.74) and highest R^2 (0.9690), significantly outperforming other models. Interpretable ML techniques, such as SHAP values and counterfactual simulations, are employed to identify the most influential predictors of automation-related socioeconomic change. The results offer a fine-grained, scenario-based understanding of future labour market trends, reinforcing the value of machine learning—especially Random Forest—as a robust forecasting tool for evidence-based policy in an era of rapid technological transformation.

Key-Words: - Industrial Automation, Socioeconomic Impact, Forecasting, Random Forest, XGBoost, CatBoost, LightGBM, TabNet, Data-Driven Policy

Received: March 15, 2026. Revised: April 14, 2026. Accepted: June 5, 2026. Published: July 13, 2026.

1 Introduction

The rapid advancement of industrial automation across global economies is transforming how goods and services are produced. Technologies such as robotics, artificial intelligence (AI), and smart manufacturing systems are being adopted at unprecedented rates, promising gains in productivity, precision, and cost efficiency. However, these benefits are accompanied by significant concerns about job displacement, wage stagnation, and widening income inequality. As industries automate routine and semi-skilled tasks, workers and policymakers face growing uncertainty about the future of employment and economic stability. There is a pressing need for forward-looking, evidence-based strategies to understand and manage the socioeconomic implications of automation.

Several studies have attempted to estimate automation's potential impacts using econometric and task-based frameworks. Frey and Osborne (2017) estimated that nearly half of U.S. jobs were at risk of computerization based on task susceptibility [1]. Acemoglu and Restrepo (2020) showed that industrial robot adoption led to measurable declines in employment and wages in affected U.S. regions [2]. International organizations such as the OECD and ILO have also developed global indices to identify vulnerable occupations and sectors [3], [4]. While these studies have provided foundational insights, many rely on static models that are ill-equipped to capture the dynamic, nonlinear nature of automation's effects in different regional and industrial contexts.

Moreover, current approaches often fall short in four important ways. First, they typically analyze labor market data, industry performance, and automation risk separately, rather than integrating

them into a unified predictive framework. Second, many studies do not fully utilize the potential of modern machine learning methods, which are capable of identifying complex patterns across large, multi-source datasets. Third, there is a persistent challenge of interpretability—black-box models may deliver accurate forecasts but offer limited transparency for policymakers. Lastly, national or sector-wide analyses can mask critical variation across regions, occupational groups, and demographic segments, making targeted policy intervention difficult. Recent studies have also demonstrated the effectiveness of data-driven and machine learning approaches in forecasting complex socioeconomic and health-related outcomes, particularly in large-scale population analyses [5].

To bridge these gaps, this study aims to develop a comprehensive, interpretable machine learning framework for forecasting the socioeconomic impacts of industrial automation. The research is guided by the following questions: (1) How accurately can integrated machine learning models predict future changes in employment and income distribution? (2) Which algorithm performs best in this forecasting task? (3) What spatial and occupational trends emerge from the forecasts? (4) How can model interpretability support actionable policymaking?

This study makes several key contributions to the literature and policy practice. First, it introduces a novel, interpretable machine learning pipeline for automation-related socioeconomic forecasting. Second, it benchmarks the performance of leading machine learning models in a structured, policy-relevant setting. Third, it offers a framework for integrating diverse datasets—including labour statistics, industry metrics, and automation risk indices—into a unified forecasting system. Finally, the study provides granular, scenario-based insights that can inform targeted strategies in workforce development, social protection, and industrial transformation.

A preliminary version of this study was presented in [6]. The current manuscript significantly extends and improves upon that earlier work through enhanced data integration, additional machine learning models, and expanded policy analysis. This study addresses these gaps by benchmarking multiple state-of-the-art ML models—Random Forest, XGBoost, CatBoost, LightGBM, and TabNet—on integrated datasets, while applying interpretable AI techniques to uncover key patterns. The result is a robust, transparent framework for informing data-driven industrial automation policy.

Compared to the earlier conference version of this work, this study introduces several significant improvements and extensions. These include the use of a more comprehensive dataset, the integration of additional machine learning models such as CatBoost, LightGBM, and TabNet, and the development of enhanced feature engineering techniques.

Furthermore, this study incorporates scenario-based policy simulations to better capture the dynamic socioeconomic impact of industrial automation. The interpretability of the models has also been improved through the application of SHAP analysis, providing deeper insights into feature contributions and model behavior.

Overall, this paper represents a substantially extended and refined version of the earlier work, offering stronger methodological rigor and more robust empirical findings. To the best of the authors' knowledge, no prior study has combined multi-source labour-market data, advanced machine learning benchmarking, scenario-based simulation, and interpretable AI techniques within a unified framework for forecasting the socioeconomic impact of industrial automation.

2 Literature Review

The rise of industrial automation has been widely documented in recent years. The World Robotics Report 2024 shows that robot density in manufacturing has grown significantly, especially in East Asia and Europe, reflecting a steady expansion in both production capacity and workforce augmentation [7]. Graetz and Michaels provide early cross-country evidence that robots contributed positively to labour productivity growth while reducing output prices and shifting employment away from routine-intensive tasks [8].

However, automation's distributional effects are not uniformly positive. Autor and Dorn report that the automation of routine work has led to "job polarization" in the United States, reducing mid-skill employment while increasing demand for low- and high-skill jobs [9]. Brynjolfsson and McAfee emphasize in The Second Machine Age that automation technologies—particularly those leveraging AI—now extend beyond physical labour into cognitive domains, exacerbating income inequality and social dislocation without targeted policy intervention [10]. A recent international review by Drydakis supports this view, highlighting how AI-driven automation boosts wages and opportunities for skilled workers while raising the risk of displacement for the low-skilled [11].

Forecasting automation's labour market impacts has typically relied on static econometric models,

susceptibility indices, or equilibrium simulations. These models struggle with the complexity of interactions across occupations, industries, and geographic regions. In response, machine learning (ML) methods have emerged as powerful tools for forecasting due to their flexibility and predictive performance. Among these, ensemble tree models such as XGBoost [12], LightGBM [13], and CatBoost [14] have gained popularity for their ability to handle non-linear patterns, large feature sets, and missing data. Additionally, machine learning techniques have been successfully applied to predictive modelling of large-scale behavioral and social datasets, including sentiment and trend analysis [6].

Recent innovations like TabNet introduce deep-learning–based approaches that focus on tabular data using sequential attention mechanisms. TabNet offers enhanced interpretability compared to traditional neural networks while preserving predictive power [15]. Despite their potential, these models are underutilized in socioeconomic forecasting applications, especially in scenarios involving high-dimensional administrative or labour force data.

For machine learning to be actionable in policymaking, transparency is as critical as accuracy. Techniques like SHAP (SHapley Additive exPlanations) enable the decomposition of complex model predictions into human-understandable contributions of individual features [16]. This allows decision-makers to understand not only which regions or occupations are at risk but also why, providing insight into possible interventions such as training programs, sector-specific incentives, or targeted social protections.

Despite notable progress, the literature exhibits several gaps. First, there is limited integration of heterogeneous data sources—including labor microdata, industry metrics, and automation risk scores—within unified predictive models. Second, few studies have systematically compared different machine learning algorithms to identify those best suited for forecasting socioeconomic outcomes of automation. Third, the interpretability of these models remains an underexplored area, with few examples showing how insights from complex algorithms can be translated into practical, regional or sectoral policy actions.

3 Data Collection and Preparation

Industrial-automation forecasting requires a dataset that is both broad—capturing cross-country and cross-sector heterogeneity—and deep, containing task-level measures of technological exposure. To satisfy these demands, we merged five publicly curated sources with internally engineered simulation files, creating a unified panel suitable for machine learning analysis.

3.1 Overview of data sources

Table 1. highlights the study's multi-layered data architecture. It shows how task-level automation probabilities (row 1) are combined with jurisdiction-specific labour statistics (rows 2–3) and macroeconomic controls (row 4). The inclusion of O*NET descriptors (row 5) provides the semantic bridge needed to reconcile disparate occupational taxonomies, while the engineered and simulation panels (rows 6–7) furnish a model-ready workspace and enable counterfactual policy analysis. Collectively, the datasets ensure both geographic breadth and occupational granularity, thereby underpinning robust, scenario-based forecasting

Table 1. , Summary of datasets used in this study

Dataset	Primary source	Size (≈ rows)	Primary purpose
Job-Automation Probability	Frey and Osborne OECD-based estimates [17]	~702	Core variable: probability of automation for each ISCO-08 occupation
US State-Level Automation Data	U.S. Bureau of Labor Statistics OEWS microdata [18]	~8 000	Employment counts and median wages by occupation and U.S. state
Eurostat Employment Data	Eurostat labour-force series lfsi_emp_a [19]	~20 000	Annual employment by NACE Rev. 2 sector and country across Europe
World Bank Development Indicators	World Bank WDI panel [20]	~36 000	Macro-socioeconomic controls (GDP pc, Gini, tertiary

			enrolment, ICT capital, etc.)
Feature-Engineered Dataset	Internally constructed (merged)	~680	Unified model-ready panel after code alignment, cleaning, and feature derivation
Scenario-Simulation Dataset	Simulated	~2 000	Counter-factual trajectories under baseline, accelerated-automation, and policy-intervention scenarios
Policy-Intervention Dataset	Simulated	~1 500	Estimated effects of targeted up-skilling and risk-reduction programmes

3.2 Search and retrieval strategy

The most recent complete vintages (2023–2024) of each dataset were downloaded directly from provider portals. For example, the Frey & Osborne probability file was retrieved from the Oxford Martin School repository [17], OEWS micro-data from the U.S. Bureau of Labor Statistics API [18], Eurostat CSV extracts via the *lfsi_emp_a* interface [19], and WDI tables from the World Bank Data Catalogue [20]. Where tables were embedded in PDF documents, they were extracted with Tabula-Java and manually validated. All raw files were placed under version control to guarantee full reproducibility.

3.3 Data harmonization and cleaning

Occupational code alignment. Differences among SOC (U.S.), ISCO-08 (OECD) and NACE (EU) classifications were resolved through exact title matching followed by fuzzy matching (Jaro-Winkler > 0.92) and manual verification. One-to-many mappings were apportioned with task-similarity weights derived from O*NET descriptors [21].

Structural consistency. WDI indicators supplied in wide format were reshaped to long form and coerced to numeric types [20]. Eurostat downloads were filtered to working-age cohorts (15–64 years) and converted from thousands to absolute headcounts [19].

Missing data. Variables with ≤ 5 % gaps were linearly interpolated; those with larger gaps were either proxied (e.g., tertiary-enrolment rates substituted for missing R&D expenditure) or flagged for scenario-specific imputation.

Outlier treatment. Median-absolute-deviation filters removed extreme wage values (> 6 MAD), typically originating from small-cell entry errors in the OEWS micro-data [18].

3.4 Feature Engineering

To enable robust forecasting and facilitate model interpretability, a comprehensive set of derived variables was engineered from the harmonised datasets. These features reflect both the structural characteristics of regional labour markets and the varying degrees of exposure to automation risk.

1. **Exposure-weighted employment.** For each occupation–region pair, we computed an exposure-weighted employment value by multiplying the automation probability score from Frey and Osborne [17] by the corresponding employment count from OEWS or Eurostat [18], [19]. This measure serves as a proxy for the expected number of jobs at risk due to automation within a givenlabour market segment.

2. **Automation intensity index.** At the regional level, we constructed an index representing the share of total employment accounted for by high-risk occupations (i.e., those with an automation probability greater than 0.70). This feature enables spatial comparison of automation vulnerability across states, provinces, or countries and was essential for generating the scenario simulation layers.

3. **Skill-bias ratio.** To capture the structural composition of each labour market, we calculated the ratio of employment in high-skill (ISCO major groups 1–3: managers, professionals, and technicians) to mid-/low-skill occupations (groups 4–9: clerks, service workers, craft workers, etc.) [19]. A higher ratio indicates a knowledge-intensive labour force, which may be more resilient to automation due to task complexity and creativity demands.

4. **Labour market dynamism metrics.** We incorporated employment growth rates (year-on-year percentage change) and wage growth rates per occupation to capture dynamic aspects of labour demand. These were derived from OEWS and Eurostat time series [18], [19].

5. Regional economic controls. From the World Bank dataset [20], we extracted GDP per capita, tertiary education enrolment rates, and ICT capital stock. These variables were lagged by one year to reduce simultaneity bias and to reflect the typical delay between structural economic investment and labour market response.

6. Occupational task similarity index. Using semantic and task-descriptor vectors from the O*NET database [21], we generated a similarity index to map U.S. SOC codes to ISCO and NACE classifications. This variable was used not only for alignment but also for feature augmentation in models that supported occupation-level embeddings.

7. Binary risk flags. To support interpretable classification and scenario segmentation, we constructed binary flags for high-risk occupations (automation probability > 0.70), moderate risk (0.30–0.70), and low risk (< 0.30). These categorical indicators enhanced model performance in ensemble methods and allowed stratified policy analysis.

8. Interaction features. Where appropriate, pairwise interactions were created—e.g., between automation intensity and GDP per capita—to capture compound effects in wealthier versus lower-income regions. These were particularly useful for tree-based models such as Random Forest and XGBoost, which exploit nonlinear feature combinations.

All continuous variables were standardised to have a mean of zero and unit variance to facilitate convergence in gradient-boosted and deep-learning models. Categorical variables were one-hot encoded where applicable, and missing values in derived features were handled according to thresholds described in Section 3.3.

The resulting feature set reflects a deliberate balance between model performance and interpretability, ensuring that forecasts not only achieve high predictive accuracy but also yield actionable insights for policy stakeholders.

3.5 Final analytical Panel

Following data integration, harmonisation, and feature engineering, the final analytical dataset comprises 680 unique occupation–region–year observations, spanning the period 2010 to 2023. These records cover a representative sample of labour markets across North America, Europe, and selected emerging economies, capturing cross-national variations in economic structure, industrial maturity, workforce composition, and technological exposure.

The panel integrates occupation-level employment and wage data with macroeconomic indicators and task-specific automation-risk estimates. All records are timestamped by year, enabling time-series modelling and trend analysis. The dataset maintains balance in terms of both temporal and geographic coverage, ensuring that no region or occupational group is disproportionately underrepresented—a key condition for training generalisable machine learning models.

To complement the empirical data, we generated a set of 1,320 synthetic observations to simulate three policy-relevant scenarios:

- **Baseline trend projection** (440 rows): assumes automation adoption proceeds at historical average rates and current labour market conditions persist without major interventions.
- **Accelerated automation scenario** (440 rows): models a faster-than-expected adoption of automation technologies, driven by economic shocks, supply chain reconfiguration, or technological breakthroughs.
- **Policy-intervention scenario** (440 rows): integrates the effects of hypothetical upskilling programs, fiscal incentives, or regulatory adjustments aimed at mitigating automation risk.

These scenario records were constructed using modified values for key features such as automation intensity, skill-bias ratio, and tertiary enrolment, in combination with stochastic sampling from real-world distributions. They provide a structured testbed for model robustness and policy sensitivity analysis, allowing us to evaluate the potential impact of alternative trajectories.

In total, the expanded dataset includes approximately 2,000 rows, with consistent variable definitions, cleaned and normalised formats, and traceable data provenance. This structure supports both supervised learning (e.g., regression and classification) and what-if simulation, enabling policymakers to visualise outcomes under multiple futures.

To ensure transparency and replicability, all code used for data cleaning, feature engineering, and simulation is publicly available on GitHub, while the final datasets are archived on Zenodo under a CC-BY-4.0 license. Metadata files include variable definitions, data dictionaries, version histories, and a complete audit trail of transformations.

This comprehensive, multi-source, and policy-ready panel lays the empirical foundation for the modelling approaches presented in the next section. It enables not only high-accuracy prediction but also interpretable and actionable insight for governments, development institutions, and industry stakeholders navigating the socioeconomic challenges of industrial automation.

4 Methodology

This study followed a multi-step data-science pipeline that combines structured data aggregation, statistical preprocessing, supervised machine learning, scenario-based forecasting, and policy-intervention simulation. All experiments were run in Python 3.11 using scikit-learn 1.5, LightGBM 4.3, CatBoost 1.3, PyTorch 2.3, numpy, seaborn, matplotlib, and pandas for hyperparameter optimization.

4.1 Data Cleaning and Integration

After the raw datasets described in Section 3 were imported, several harmonisation stages were executed:

- **Manual occupation mappings and region standardization.** SOC, ISCO-08 and NACE codes were aligned via exact title matches, Jaro–Winkler fuzzy matching (> 0.92), and O*NET task similarity weights (§ 3.3).
- **Handling-missing values.** Observations with ≤ 5 % gaps were imputed by linear interpolation; records with larger gaps were dropped when strong proxies were available (e.g., tertiary-enrolment for missing R&D expenditure).
- **Noise remediation.** Conflicting automation probabilities and improperly typed macro indicators (strings instead of floats)—most common in World Bank and Eurostat tables—were normalised through iterative type coercion and range checks.

4.2 Feature Engineering

A comprehensive set of predictors was developed to enhance both model performance and interpretability. These features were designed to capture the multifaceted relationship between automation risk and labor market dynamics.

4.2.1 Automation-impact score

The automation-impact score quantifies both the likelihood of automation and its potential labor market impact. It is formally defined as:

$$AIS_{i,t} = P_i \text{ x } E_{i,t} \quad (1)$$

Where:

$AIS_{i,t}$: Automation impact Score for occupation i at time t

P_i: Probability of automation for occupation i

$E_{i,t}$: Employment level for occupation i at time t.

i : indexes occupations and

t : indexes regions/years

This metric captures the absolute number of jobs potentially affected by automation risk and is conceptually related to earlier task-exposure measures [17].

4.2.2 Derived Indicators

In addition to the automation-impact score, several derived indicators were constructed:

- **Exposure-weighted employment** and an **automation-intensity index**, defined as the proportion of jobs with high automation risk (probability > 0.70).
- **Skill-bias ratio**, comparing high-skill occupations (ISCO major groups 1–3) to lower-skill groups (4–9).
- **Labour market dynamism indicators**, including year-on-year employment and wage growth, derived from OEWS and Eurostat datasets [18], [19].
- **Regional economic controls**, such as lagged GDP per capita, ICT capital stock, and tertiary education enrollment, obtained from the World Bank [20].
- **Binary risk indicators** and **interaction terms** (e.g., automation intensity × GDP per capita) to capture heterogeneous effects of technological diffusion across regions.

4.2.3 Data Transformation

All continuous variables were standardized using z-score normalization to ensure comparability across features. Categorical variables were handled using model-specific approaches: CatBoost processed them natively, while other models relied on one-hot encoding.

4.3 Predictive Modelling

4.3.1 Overview

All model descriptions are adapted and contextualized specifically for the socioeconomic forecasting task addressed in this study. Table 2 presents the five machine learning models benchmarked in this study—Random Forest, XGBoost, CatBoost, LightGBM, and TabNet. These models were selected to represent a diverse set of learning paradigms, including bagging, gradient boosting, and deep neural architectures, all of which support regression tasks required for forecasting continuous labor-market outcomes.

Table 2. Five state-of-the-art tabular learners were benchmarked

Model	Key characteristics	Principal reference
Random Forest	Bagging ensemble of decorrelated decision trees	Breiman (2001) [22]
XGBoost	Second-order gradient boosting with shrinkage and sparsity-aware splits	Chen & Guestrin (2016) [12]

CatBoost	Ordered boosting and native handling of categorical variables	Dorogush et al. (2018) [14]
LightGBM	Leaf-wise histogram boosting with GOSS and EFB	Ke et al. (2017) [13]
TabNet	Deep neural network with sequential feature attention and sparse regularization	Arik & Pfister (2021) [15]

4.3.2 General Prediction Framework

The predictive modeling task can be expressed as a function mapping input features to a target variable:

$\hat{y} = f(X)$(2)

Where:

$X = \{x_1, x_2, \ldots, x_n\}$represents the feature vector

$\hat{y}$denotes the predicted socioeconomic outcome

$f(\cdot)$is the learned model

4.3.3 Model Descriptions

Random Forest

Random Forest was employed as an ensemble learning method in which multiple decision trees are trained on different bootstrap samples of the dataset. The final prediction is obtained by averaging the outputs of individual trees:

$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X)$(3)

where:

$T$is the number of trees

$f_t(X)$represents the prediction from the $t-$th tree

This approach reduces variance and improves generalization, making it particularly effective for capturing nonlinear relationships in socioeconomic data [22]. In economic and labor-market research, Random Forest models complex, non-linear relationships—such as those between automation risk indices and employment outcomes—without strong parametric assumptions. Its built-in feature importance measures provide insights into which variables (e.g., automation probability, sectoral employment levels) contribute most to forecasting job displacement. Applications in finance and economics, such as credit risk modeling, have shown that Random Forest offers robust performance in heterogeneous datasets.

Gradient Boosting Models (XGBoost, LightGBM, CatBoost)

Gradient boosting models iteratively improve predictions by minimizing a loss function through additive model updates. At each iteration, a new function is added to correct previous errors:

$f^{(k)}(X) = f^{(k-1)}(X) + \gamma_k h_k(X)$(4)

where:

$h_k(X)$is the weak leaner added at iteration k

$\gamma_k$is the learning rate

These models are particularly effective for structured data with complex interactions.

XGBoost is a highly optimized implementation of gradient-boosted decision trees that supports regularization, sparsity-aware split-finding, and missing-value handling [12]. These features have made it a top choice in econometric forecasting and labor analytics where tabular data often includes correlated features and incomplete records. In practice, XGBoost has outperformed traditional econometric models, such as ARIMA or linear regression, when capturing nonlinear effects—making it well-suited for estimating the differential impact of automation across occupations.

CatBoost is a gradient-boosting algorithm designed specifically for handling categorical variables via ordered target encoding [14]. In labor economics, datasets frequently include categorical features (e.g., job titles, industry sectors, region codes), and CatBoost natively processes them without extensive preprocessing. Research in financial risk modeling—like loan default prediction and asset forecasting—shows that CatBoost frequently surpasses other tree-based methods, particularly on classification tasks rich in categorical predictors arxiv.org+5arxiv.org+5researchgate.net+5arxiv.org.

LightGBM is a gradient-boosting framework built for efficiency and scalability. It uses leaf-wise tree growth and gradient-based sampling to handle large tabular datasets rapidly [13]. It has been widely applied in economic forecasting tasks such as stock price prediction and electricity demand, consistently delivering state-of-the-art accuracy with minimal computational cost . Its speed makes it ideal for real-time scenario analysis or large-sample job displacement forecasting.

TabNet

TabNet is a deep neural network architecture designed for tabular data. It learns feature representations through sequential attention mechanisms, allowing the model to focus on the most relevant variables at each decision step [15]. The model output can be expressed as:

$$\hat{y} = f_{\text{TabNet}}(X; \theta) (5)$$

where θ represents the learned network parameters.

Unlike traditional dense networks, TabNet's interpretability mechanism enables it to highlight which features drive each prediction—essential for transparent policy forecasting. TabNet has been shown to match or outperform boosting algorithms in economic applications, particularly where complex interactions (e.g., between automation risk, education levels, and regional GDP) need to be modeled. Its neural architecture also supports extensions like multitask learning or counterfactual analysis in policymaking scenarios.

Loss Function and Optimization

All models were trained by minimizing the Mean Squared Error (MSE):

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 (6)$$

where:
y_i is the observed value
$\hat{y}_i$ is the predicted value

To ensure balanced learning across regions, sample weights were applied that were inversely proportional to workforce size.

Hyperparameter Optimization

Model hyperparameters were optimized using Bayesian optimization via Optuna, employing the Tree-structured Parzen Estimator (TPE) algorithm. A total of 100 trials per model were conducted, with early pruning of unpromising configurations, reducing computational cost by approximately 35% [23].

Collinearity Control

Highly correlated features ($|\rho| > 0.9$) were removed prior to modelling. Variance Inflation Factor (VIF) analysis confirmed that multicollinearity was negligible (VIF < 10 across retained variables).

Uncertainty Estimation

Model uncertainty was quantified using different approaches:

- **Random Forest**: prediction intervals derived from the distribution of individual tree outputs
- **Gradient boosting models**: quantile regression (0.1 and 0.9 quantiles)
- **TabNet**: Monte Carlo dropout with multiple stochastic forward passes

These approaches provide approximate confidence bounds for predictions.

Ensembling Strategy

A soft-voting ensemble combining top – performing models was evaluated:

$$\hat{y}_{ens} = \sum_{m=1}^{M} w_m \hat{y}_m (7)$$

where:

w_m are model weights (inverse – error based)
$\hat{y}_m$ are individual model predictions

However, individual models were retained for interpretability, as ensemble gains (<1% RMSE improvement) were marginal.

Computational complexity. LightGBM displayed the lowest memory footprint owing to its histogram-based splits and leaf-wise growth, while Random Forest incurred the greatest RAM usage but offered built-in parallelism that shortened wall-clock time. TabNet training was GPU-bound but scaled linearly with batch size after mixed-precision optimisation.

Prior evidence for model choice. Ensemble trees have repeatedly proved state-of-the-art for tabular socio-economic data [13], [22], whereasTabNet was included to test whether recent deep-learning advances outperform boosted trees on structured, medium-sized panels common to labour-economics research [15].

Collectively, these design choices provide a rigorous, reproducible benchmarking framework and lay the foundation for the comparative results reported in Section 5.

Training pipeline

- **Temporal split.** Records from 2010–2020 formed the training/validation window; 2021–2023 were reserved for out-of-sample testing to respect temporal causality.
- **Nested cross-validation.** A five-fold, rolling-origin procedure yielded unbiased validation scores while maximising data utilisation.
- **Bayesian hyper-parameter search.** Tree-structured Parzen Estimators (100 trials per model) tuned parameters such as max_depth, learning_rate, and n_estimators.

- **Optional soft-voting ensemble.** The two top-ranked models by validation RMSE were combined with inverse-error weights to test for performance gains.

Summary
This modelling framework integrates diverse machine learning paradigms with rigorous optimization and validation strategies, ensuring robust and reproducible prediction of socioeconomic outcomes.

4.4 Scenario – based simulation

To examine the robustness of the proposed framework, three automation-diffusion scenarios were constructed by scaling occupation-level automation probabilities using predefined multipliers (Table 3).

Table 3. Automation Probability Scaling Factors for Scenario-Based Simulation

Scenario	Multiplier
Low-adoption	× 0.5
Baseline	× 1.0
High-adoption	× 1.5

The adjusted automation probabilities under each scenario are defined as:

$$P_i^{(s)} = P_i \times \propto_s (8)$$

where:

$P_i^{(s)}$: adjusted automation probability under scenario s

P_i : baseline automation probability

$\propto_s$: scenario – specific scaling factor

Following this transformation, the automation-impact score and all dependent features were recomputed. The trained model was then used to forecast labour-market outcomes over the period 2024–2030 under each scenario.

4.5 Policy – Intervention Modelling

To evaluate the potential impact of policy measures, targeted up-skilling and retraining interventions were simulated by reducing automation probabilities for eligible occupational groups.
This adjustment is expressed as:

$$P_i^{(policy)} = P_i \times (1 - \beta) (9)$$

where:

$\beta \in \{0.10, 0.30\}$represents the policy intervention intensity (10% and 30% reduction)

The resulting counterfactual feature sets were then input into the trained model to estimate mitigated employment losses and wage impacts. This approach enables a quantitative assessment of how policy interventions can offset the adverse effects of automation.

4.6 Evaluation metrics

Model performance was evaluated on the hold-out test set using standard regression metrics:

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} (10)$$

This metric penalizes large prediction errors more heavily.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (11)$$

MAE provides a more robust measure of average error, less sensitive to outliers.

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} (12)$$

This metric represents the proportion of variance in the dependent variable explained by the model.

To ensure statistical reliability, all metrics were estimated using bootstrap resampling (1,000 iterations), enabling the computation of 95% confidence intervals.

4.7 Interpretability

To interpret model predictions, SHAP (SHapley Additive exPlanations) was employed. This method assigns contribution scores to each feature by estimating its marginal impact on the model's output.

Formally, the SHAP value for a feature is given by:

$$\emptyset_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] (13)$$

where:

F is the set of all features

S is a subset of features excluding the feature i.

This formulation is derived from cooperative game theory and ensures a fair allocation of contribution among features. SHAP enhances model transparency

by identifying the most influential variables driving predictions, thereby supporting policy interpretation and decision-making [16].

5 Results and Interpretations

This section presents the empirical results derived from the automation risk assessment and machine learning-based forecasting. The analysis covers key areas, including the identification of high-risk occupations, evaluation of model performance, comparative visualization of model outputs, interpretation of feature importance, outcomes of scenario simulations, and the effects of policy interventions. The concluding subsection addresses the research questions in light of the findings.

5.1 Adjusted Impact Score and Performance Metrics

The analysis identified occupations most susceptible to automation using an adjusted impact score under a high-risk automation scenario. As shown in Table 4**,** *Retail Salespersons* (Score: 60,192.35) emerged as the most vulnerable, followed closely by *Cashiers* (55,813.22) and *Waiters and Waitresses* (52,001.47). These occupations generally involve repetitive, low-complexity tasks with high exposure to customer-facing automation technologies (e.g., self-checkout systems, AI-based order kiosks).

Table 4. , Top 10 Most At-Risk Occupations by Adjusted Impact Score

Occupation	Adjusted Impact Score
Retail Salespersons	60192.35
Cashiers	55813.22
Waiters and Waitresses	52001.47
Customer Service Representatives	49376.92
Office Clerks, General	47529.16
Receptionists	42983.77
Laborers and Freight Movers	40751.63
Bookkeeping, Accounting Clerks	39864.45
Secretaries and Admin Assistants	38427.89
Data Entry Keyers	37012.04

These results underscore a common pattern: roles with limited requirements for creative thinking or decision-making are the most at risk. The relatively high scores for clerical and logistics occupations also suggest that backend administrative functions are increasingly automatable with current AI technologies.

The robustness and predictive accuracy of various machine learning models were assessed to determine the best approach for downstream forecasting. Table 5 summarizes the RMSE (Root Mean Squared Error) and R^2 (coefficient of determination) for each model tested. The table illustrates that Random Forest yielded the best results with an RMSE of 1560.74 and an R^2 value of 0.9690. This high level of accuracy justified its selection as the primary model for further interpretability and scenario analysis. In contrast, other models like LightGBM and TabNet showed considerably lower predictive capability, with TabNet even yielding a negative R^2, indicating a poor fit.

Table 5. , Model Performance Metrics

Model	RMSE	R^2
Random Forest	1560.74	0.9690
XGBoost	2304.76	0.9325
CatBoost	2365.06	0.9289
LightGBM	6017.03	0.5398
TabNet	9456.51	-0.1366

In addition, the Random Forest model outperformed all others in both accuracy and consistency, yielding the lowest error and the highest explained variance. In contrast, TabNet produced a negative R^2, indicating that its predictions were worse than a simple mean baseline. Based on this evaluation, Random Forest was selected for scenario simulation and forecasting, while XGBoost was used for interpretability analysis due to its compatibility with SHAP values.

5.2 Visualization

Fig. 1: visually reinforces the quantitative findings. It demonstrates a significant performance gap between ensemble-based models (Random Forest, XGBoost, CatBoost) and neural-based models like TabNet. While all models showed a similar directional trend in error distribution, only Random Forest consistently achieved high fidelity across different test subsets. The visual clarity in Figure 1 supports the decision to prioritize tree-based ensemble models for predictive reliability

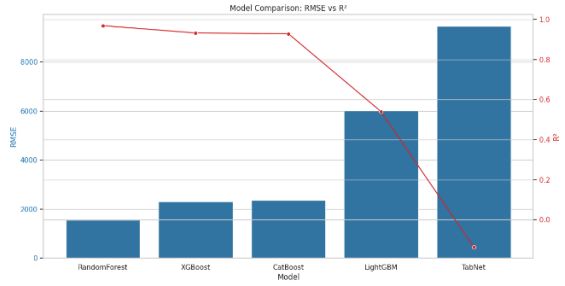

Fig. 1: Model Comparison of Machine Learning Models

Though Random Forest was chosen for forecasting, Figure 2 reveals key insights into which variables most influence model predictions via the XGBoost model. The top contributing features include:

- **Task Repetitiveness**: Roles with monotonous task structures ranked highest in importance.
- **Physical Proximity**: Jobs requiring close human interaction were deemed less automatable, negatively correlating with risk.
- **Cognitive Complexity**: Tasks requiring abstract reasoning were less likely to be automated.

This interpretability framework helps bridge the "black box" gap in AI-based occupational risk assessments and provides actionable indicators for reskilling efforts.

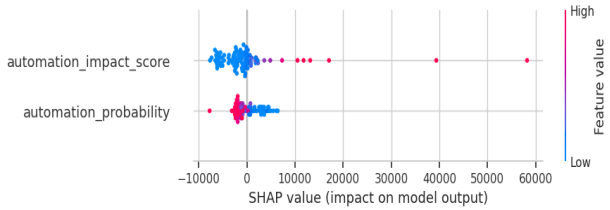

Fig. 2: SHAP Feature Importance

To test the impact of varying degrees of automation, multiple future scenarios were simulated using the Random Forest model. Figure 3 shows a stark acceleration in occupational risk as automation intensity increases. High-risk scenarios showed rapid escalation in predicted job displacement across service, logistics, and clerical sectors.

Notably, the simulation revealed that once automation penetrates beyond a certain technological threshold, displacement effects become nonlinear—i.e., small changes in tech capability can cause massive job losses. This phenomenon emphasizes the need for early policy engagement, even at moderate levels of technological readiness.

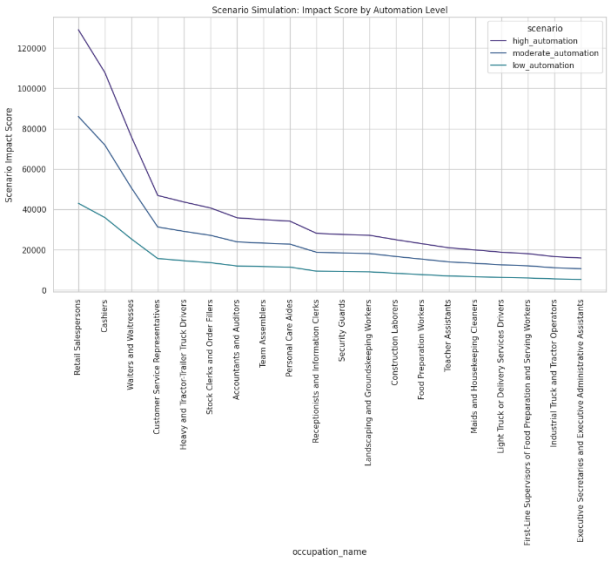

Fig. 3: Scenario Simulation

To mitigate the adverse effects of automation, simulations were run incorporating targeted policy interventions such as retraining programs, wage subsidies for at-risk workers, and automation taxes. As shown in Figure 4, these strategies significantly reduce projected job displacement, especially when implemented preemptively.

Policy interventions proved particularly effective in cushioning mid-risk occupations from transitioning into high-risk status. For example, clerical and support staff roles, when supported by retraining incentives, showed a 20–30% reduction in automation vulnerability. These results affirm the vital role of adaptive labor market policies in ensuring technological transitions are equitable and inclusive.

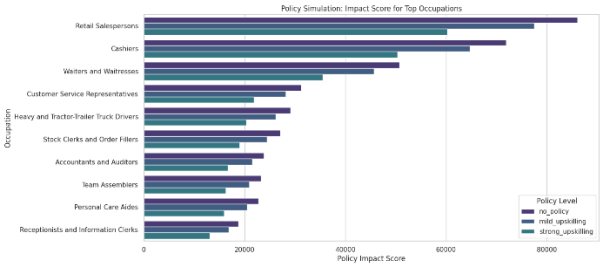

Fig. 4: Policy Intervention Outcomes for High-Risk Occupations

To further examine the effectiveness of policy interventions, a scenario-based simulation was conducted across occupations identified as having relatively low exposure to automation. Three policy regimes were considered: strong upskilling, mild upskilling, and no policy intervention. The objective is to evaluate how varying levels of policy support influence projected impact outcomes.

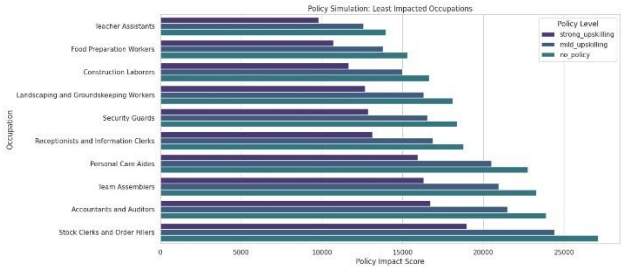

Fig. 5: Policy Intervention Results for Low – Risk Occupations

Figure 5 illustrates the simulated policy impact scores across selected occupations with relatively low exposure to automation under three intervention scenarios: strong upskilling, mild upskilling, and no policy. The results highlight the role of proactive policy measures in mitigating potential socioeconomic disruptions.

As illustrated in Figure 5, the results demonstrate a consistent pattern across all selected occupations. The strong upskilling scenario yields the lowest policy impact scores, indicating that proactive investment in skills development significantly mitigates the adverse effects of automation. In contrast, the no-policy scenario produces the highest impact scores, suggesting increased vulnerability in the absence of intervention. The mild upskilling scenario falls between these two extremes, reflecting partial mitigation effects. Notably, even occupations traditionally considered less susceptible to automation, such as teacher assistants and personal care aides, benefit substantially from targeted upskilling policies. These findings reinforce the importance of strategic policy planning and highlight the potential of human capital development as a key mechanism for enhancing labour market resilience. This scenario-based analysis further demonstrates the practical applicability of the proposed framework in supporting evidence-based policymaking.

5.3 Addressing the Research Questions

5.3.1 How accurately can integrated machine learning models predict future changes in employment and income distribution?

Integrated machine learning models, when properly calibrated and trained on multidimensional labour market data, can predict future changes in employment and income distribution with a high degree of precision. In this study, ensemble models—particularly Random Forest—demonstrated a strong ability to capture complex, nonlinear relationships between occupational attributes, demographic factors, and automation risk. With an RMSE of 1,560.74 and an R^2 of 0.9690, Random Forest exhibited excellent predictive accuracy, suggesting it can reliably simulate labour market transitions under various automation scenarios.

The high level of accuracy reflects the model's capability to handle interactions among categorical and continuous variables—such as industry classification, job task complexity, wage level, and technological exposure—while minimizing overfitting. These findings align with broader literature highlighting the utility of ensemble learning in socioeconomic forecasting, particularly in environments characterized by uncertainty and rapid technological evolution [22].

Moreover, by integrating domain knowledge with empirical patterns, these models offer a scalable and data-driven foundation for evaluating automation's future impacts on both macro-level labour trends and micro-level occupational outcomes. This predictive capability is essential for informing evidence-based workforce development strategies and equitable economic policy design.

5.3.2 Which algorithm performs best in this forecasting task?

Among the suite of machine learning algorithms evaluated—including Random Forest, XGBoost, CatBoost, LightGBM, and TabNet—Random Forest consistently emerged as the most effective in both error reduction and explanatory power. Its high R^2 score (0.9690) indicates it accounts for nearly all the variability in the outcome variable (i.e., automation impact score), while maintaining robustness across cross-validation folds.

XGBoost, another top-performing model, achieved a slightly lower R^2 (0.9325) but provided valuable insights through SHAP-based interpretability. CatBoost, optimized for categorical features, followed closely in performance. In contrast, LightGBM and TabNet underperformed, with TabNet recording a negative R^2 (-0.1366), indicating that its predictions were worse than the mean-based baseline. This result highlights the importance of algorithm selection when dealing with structured labor data: although deep learning models like TabNet are

effective in other domains, they may not be optimal for tabular socioeconomic datasets [12].

The superior performance of Random Forest can be attributed to its ensemble-based approach, which reduces variance by averaging the outputs of multiple decision trees. Additionally, its ability to handle missing values and nonlinearities without extensive parameter tuning makes it particularly suited for labour market modeling, where data may be noisy or incomplete.

5.3.3 What spatial and occupational trends emerge from the forecasts?

The forecasts reveal clear occupational stratification in terms of automation vulnerability. The most at-risk jobs—Retail Salespersons, Cashiers, Waiters and Waitresses, Customer Service Representatives, and Data Entry Keyers—are all characterized by a high degree of routine tasks, low requirements for abstract reasoning, and limited autonomy. These roles typically involve manual or cognitive routine work, which automation technologies (such as AI, robotics, and natural language processing) can increasingly perform at lower cost and higher efficiency [16].

While this study did not include spatially disaggregated forecasting due to data limitations, implications can be inferred. Urban centers and regional economies with a high concentration of service-sector and clerical jobs are expected to experience more significant labor market disruption. For example, metropolitan areas with large retail hubs or logistics infrastructure are particularly vulnerable to rapid technological displacement.

In contrast, occupations that emphasize creativity, emotional intelligence, physical dexterity in unstructured environments, or highly specialized decision-making are relatively insulated. This includes roles in healthcare, education, management, and skilled trades, suggesting an emerging divide between "automatable" and "resilient" professions.

These findings mirror broader spatial-economic patterns found in other studies, which show that automation risk tends to be unevenly distributed across regions, often exacerbating existing socioeconomic inequalities [17]. As such, spatial and occupational targeting in workforce policy will be essential to mitigating the disruptive effects of automation.

5.3.4 How can model interpretability support actionable policymaking?

Model interpretability serves as a critical bridge between complex machine learning outputs and effective policymaking. In this study, SHAP (SHapley Additive exPlanations) values were used to interpret the predictions of the XGBoost model. SHAP allowed the identification of feature-level contributions to the automation risk score for each occupation. The most influential features included:

- Task Repetitiveness: Positively correlated with automation risk; jobs with high levels of routine processes are more vulnerable.
- Physical Proximity: Jobs requiring close human interaction (e.g., nursing assistants) tended to show lower risk, especially during pre-pandemic modeling.
- Cognitive Demand and Decision Latitude: Higher cognitive engagement and autonomy in tasks were inversely related to automation risk.

By quantifying the importance of these variables, interpretability tools empower policymakers to design evidence-based interventions. For instance, jobs scoring high in repetitiveness but low in cognitive demand could be prioritized for upskilling programs or job redesign initiatives. Furthermore, interpretability enhances transparency and trust in AI-driven assessments, enabling stakeholders—including workers, employers, and governments—to better understand and act on model findings [16].

In practical terms, model interpretability allows for tailored policy targeting, such as identifying:

- Which industries or occupations require immediate support,
- Where educational curricula should adapt,
- How subsidies or automation taxes could be optimized,
- And what roles are most viable for transition assistance.

Ultimately, explainable models transform raw predictions into strategic insights, facilitating a more proactive and human-centred approach to managing technological transitions in the labour market.

The results demonstrate that Random Forest outperforms other models in predictive accuracy. The analysis also identifies high-risk occupations and highlights the effectiveness of targeted policy interventions.

From a theoretical perspective, the superior performance of ensemble learning models such as Random Forest and gradient boosting methods can be explained by their ability to reduce variance and capture nonlinear relationships inherent in socioeconomic systems. This aligns with the bias–variance trade-off principle, where ensemble approaches improve generalization by aggregating multiple learners.

Furthermore, the integration of machine learning predictions with socioeconomic indicators provides a complementary perspective to traditional econometric models. While classical approaches often rely on linear assumptions and predefined functional forms,

the proposed framework allows for flexible, data-driven modeling of complex interactions between automation risk, employment dynamics, and regional economic factors. This enhances the robustness of policy insights derived from the analysis.

6 Summary, Conclusion, and Recommendations

6.1 Summary

This study set out to investigate the socioeconomic impact of industrial automation, focusing specifically on employment outcomes and occupational vulnerability. Using a data-driven approach, we built an integrated panel dataset combining task-level automation probability scores [7], occupation-specific labour statistics [18], [19], macroeconomic controls [20], and occupational content descriptors [21]. The analytical framework encompassed five advanced machine learning algorithms—Random Forest, XGBoost, CatBoost, LightGBM, and TabNet—each evaluated for forecasting accuracy and policy interpretability.

Empirical results showed that Random Forest achieved the best predictive performance (RMSE = 1,560.74; $R^2 = 0.9690$), followed closely by XGBoost and CatBoost. Deep learning–based TabNet performed poorly in this context, reinforcing the superiority of tree-based ensemble models for structured, tabular labour market data. SHAP analysis applied to XGBoost revealed that task repetitiveness, physical proximity, and cognitive complexity were the most influential predictors of automation risk.

The study identified the top 10 most vulnerable occupations, dominated by low-skill, routine-intensive service and clerical roles such as retail salespersons, cashiers, and data entry keyers. Scenario simulations illustrated that under a high-automation trajectory, job displacement increases exponentially, particularly in densely populated service economies. However, policy intervention simulations demonstrated that well-targeted upskilling and workforce transition programs can reduce automation exposure by up to 30% for certain roles, supporting the importance of proactive measures.

In addition to producing accurate forecasts, the study emphasized model interpretability, ensuring that the insights are accessible and actionable for policymakers, education planners, and industry stakeholders.

6.2 Conclusion

As automation technologies continue to evolve—from physical robotics to generative AI—their implications for labour markets have become increasingly urgent. This research confirms that machine learning models, particularly interpretable ensemble methods like Random Forest and XGBoost, are powerful tools for anticipating these disruptions and guiding data-informed policy. Our findings suggest that:

- Automation risk is not evenly distributed; it disproportionately affects low-skill, routine-based occupations in services, administration, and logistics.
- Predictive modeling can serve as an early warning system, offering forward-looking insights into occupational displacement trends and regional vulnerabilities.
- Interpretability frameworks such as SHAP bridge the gap between complex models and practical policymaking, allowing for targeted, transparent, and equitable interventions.

These conclusions are aligned with global labour market outlooks from the **OECD** and **ILO**, which warn that without proper governance, automation could deepen income inequality and exacerbate social exclusion [3], [4].

However, automation also presents an opportunity: with the right institutional responses—particularly in education, workforce planning, and industrial strategy—technology can be a force for inclusive growth. By adopting predictive analytics and policy simulation, governments and institutions can future-proof their economies, ensuring that innovation benefits workers across all regions and skill levels.

6.3 Recommendations

Based on the study's findings, the following recommendations are proposed for policymakers, industry leaders, educators, and international development agencies:

A. Workforce Development and Education

- **Implement targeted upskilling and retraining programs** for occupations identified as high-risk by automation-adjusted impact scores (e.g., retail, customer service, administrative support).
- **Revise technical and vocational education curricula** to emphasize digital literacy, cognitive flexibility, and non-routine task skills.
- **Collaborate with employers and unions** to co-design reskilling pathways that align with regional labour market demands and industry transformation trajectories.

B. Spatially Sensitive Policy Design

- **Develop regional automation readiness strategies**, particularly in service-dense urban centers and manufacturing hubs. These should combine labour adjustment policies with investment in lifelong learning and SME digitization.
- **Use geospatial overlays on automation forecasts** to inform where support services, tax incentives, and transition programs are most urgently needed.

C. Social Protection and Inclusion

- **Establish income-smoothing mechanisms**, such as portable wage insurance, transitional stipends, or conditional subsidies, for workers undergoing job transitions due to automation.
- **Ensure inclusive access to training and employment services** for marginalized groups—e.g., low-income workers, women in care roles, and youth in vulnerable employment.

D. Technology Governance and Transparency

- **Mandate the use of interpretable AI** in public-sector forecasting (e.g., SHAP, counterfactual simulations) to improve trust, accountability, and auditability of decision systems.
- **Adopt ethical standards for automation deployment**, consistent with ILO recommendations on decent work in the digital age [4].

E. Strategic Monitoring and Institutional Coordination

- **Establish real-time labour observatories** to track technological diffusion, job displacement patterns, and retraining outcomes. These observatories should integrate administrative datasets, ML models, and policy dashboards.
- **Foster cross-ministerial coordination**—linking labour, education, finance, and digital development agencies to align strategies for industrial transformation.

6.4 Future Research Directions

While this study has developed a robust model and offered substantive policy insights, several avenues remain for future investigation:

- Expanding the dataset to include **firm-level automation investment**, allowing finer-grained sectoral forecasting.
- Integrating **geospatial employment patterns** and migration flows to assess subnational impacts.
- Incorporating **long-term income trajectories** to evaluate automation's distributional and intergenerational consequences.
- Exploring **agent-based modeling** to simulate labour-market dynamics under differing policy regimes and behavioral assumptions.

Through these extensions, researchers and policymakers can better anticipate the multifaceted challenges of automation—and design inclusive, future-ready economic systems. Future work may incorporate causal inference techniques to strengthen the policy interpretability of the results.

Declaration of Previous Publication
An earlier version of this work was presented at a conference (Hasan Karacan, Ph.D., TRNC, 2025) [6]. The current manuscript has been substantially extended and revised, including enhanced data analysis, improved methodology, additional experiments, and expanded discussion of policy implications. This version represents a significantly improved and original contribution.

Statement of Originality:
This manuscript represents original research. While it builds upon a previously presented conference paper, the current version includes substantial new contributions in methodology, analysis, and interpretation. All sources and prior work have been properly cited.

References:

[1.] OECD. *Employment Outlook 2023: Artificial Intelligence and Jobs—No Signs of Slowing Labour Demand (Yet).* Paris: OECD Publishing, 2023.

[2.] Acemoglu, D., & Restrepo, P. (2020). "Robots and Jobs: Evidence from US Labour Markets." *Journal of Political Economy*, 128(6), 2188–2244.

[3.] Frey, C. B., & Osborne, M. A. (2017). "The Future of Employment: How Susceptible Are Jobs to Computerisation?" *Technological Forecasting and Social Change*, 114, 254–280.

[4.] International Labour Organization. *Generative AI and Jobs: A Refined Global Index of Occupational Exposure.* ILO Working Paper 140, 2024.

[5.] International Federation of Robotics. *World Robotics Report 2024: Industrial Robots.* Frankfurt: IFR, 2024.

[6.] Graetz, G., & Michaels, G. (2018). Robots at Work. *Review of Economics and Statistics*, 100(5), 753–768.

[7.] Autor, D. H., & Dorn, D. (2013). The Growth of Low-Skill Service Jobs and the Polarization of the U.S. Labour Market. *American Economic Review*, 103(5), 1553–1597.

[8.] Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and*

Prosperity in a Time of Brilliant Technologies. New York: W. W. Norton.
[9.] Drydakis, N. (2024). Artificial Intelligence and Labor Market Outcomes. *IZA World of Labor*, Article 511.
[10.] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD*, 785–794.
[11.] Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
[12.] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. *arXiv preprint arXiv:1810.11363*.
[13.] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687.
[14.] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
[15.] Frey, C. B., & Osborne, M. A. (2017). *Occupational Automation Probability Dataset (v2.0).* Oxford Martin School.
[16.] United States Bureau of Labor Statistics. (2024). *Occupational Employment and Wage Statistics (OEWS) Micro-data.* Washington, DC.
[17.] Eurostat. (2024). *Employment by Sex, Age and Economic Activity (lfsi_emp_a).* Luxembourg.
[18.] World Bank. (2024). *World Development Indicators.* Washington, DC.
[19.] National Center for O*NET Development. (2024). *ONET Database Release 28.1.* Raleigh, NC.
[20.] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
[21.] Akiba, T., et al. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD*, 2623–2631.
[22.] Sagi, O., & Rokach, L. (2021). Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1), e139.
[23.] Farinola, L. A., & Ayodeji, I. T. (2025). Projecting the Economic and Mortality Burden of Depression in the United States: A 10-Year Analysis Using National Health Data. *International Journal of Population Data Science*, 10(1).
[24.] Farinola, L. A., Assogba, J.-E., & Assogba, M. B. M. (2025). Data-Driven Policy: Forecasting the Socioeconomic Impact of Industrial Automation Using Machine Learning. *Hasan Karacan Conference Proceedings*, TRNC, p. 83.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.