DATAMACLEA'26

Econometric Research Association
3rd International Data Analytics and Machine Learning Conference

Artificial Intelligence and Economic Efficiency: Sectoral Applications and Value Generation

FULL TEXTS BOOK

ECONOMETRIC RESEARCH ASSOCIATION

ANKARA YILDIRIM BEYAZIT ÜNİVERSİTESİ AYBÜ

ULİSA Uluslararası İlişkiler ve Stratejik Araştırmalar Enstitüsü

Conference Overview

The Econometric Research Association (ERA) successfully organized the 3rd International Data Analytics and Machine Learning Conference (DATAMACLEA'26) on May 11-13, 2026. The event was held both face-to-face at Ankara Yıldırım Beyazıt University, Etlik Milli İrade Campus and online, ensuring broad international participation.

This year's main theme was: "Artificial Intelligence and Economic Efficiency: Sectoral Applications and Value Generation"

The conference emphasized how data has become central to policymakers and regulators, serving as the "fuel" of the digital economy while also emerging as a key geopolitical issue. Despite the growing importance of data, the underlying "economics of data" remains underexplored. The conference aimed to foster a deeper understanding of data's economic features, such as its nature as an economic good and the value chain associated with it. The discussions stressed the need for sound economic reasoning to guide data policy, regulation, and the design of data spaces.

The event was organized in cooperation with Ankara Yıldırım Beyazıt University, Institute for International Relations and Strategic Research (ULISA- IIRSR).

Artificial Intelligence (AI) is increasingly integrated into the core of economic, technological, and organizational systems, fundamentally altering how value is created, labor is distributed, and productivity is optimized. From AI-assisted software development pipelines to data-driven decision-making in corporate operations, intelligent systems are emerging as key drivers of efficiency and innovation across industries. Yet, these transformations prompt critical economic reflections: Which human-led tasks can be responsibly automated? What are the measurable benefits and hidden costs of delegating functions to AI? And how can institutions maximize value generation while maintaining adaptability and accountability? DATAMACLEA'26 welcomes empirically grounded and forward-looking contributions that investigate these questions through technical insight, sectoral analysis, and real-world application.

At the same time, the growing adoption of AI technologies across sectors is unlocking new forms of economic efficiency, operational resilience, and scalable innovation. From intelligent documentation and simulation systems to decision-support tools and remote service platforms, AI-powered solutions are enabling organizations to optimize processes, reduce resource waste, and enhance outcome quality. These tools not only assist professionals in managing complex tasks but also generate measurable value at the institutional, national, and global levels. DATAMACLEA'26 encourages contributions that examine how such AI-enabled innovations reshape economic models, improve productivity, and drive value creation across diverse domains, including healthcare, software, education, finance, and beyond.

Topics Covered

- Artificial Intelligence in Economics / Finance / Management
- Policy, Regulation and Social Impact of Artificial Intelligence
- Machine Learning / Deep Learning
- Data Mining / Big Data
- Emerging Technologies (IoT, Blockchain, Digital Twins)
- Natural Language Processing / Financial Data Analytics
- Predictive Analytics / Forecasting
- Statistical Learning and Inference
- Econometric Theory
- Applied Economics
- Behavioral and Experimental Economics
- Agent Based Modelling

- Optimization / Operations Research
- Other

ISBN	978-605-4929-49-8
Editor(s)	Sıdıka Başcı, İbrahim Demir, Muhammed Oruç, Elmas Bener Otacı
Publisher	Ankara Yıldırım Beyazıt University
E-mail	info@era.org.tr
Webpage	www.era.org.tr/datamaclea26/
Conference Dates	11-13 May, 2026
Conference Address	Ayvalı Mahallesi, Gazze Caddesi, No.7 06010 Etlik, Keçiören, Ankara, Türkiye.
Shipping Address	Beytepe Mahallesi, Piri Reis Caddesi, 5360 Sokak, 2/2, Çankaya, Ankara, Türkiye.

May 2026

Honorary Committee

Ali Cengiz KOSEOGLU, Ph.D. Prof., Rector, Ankara Yıldırım Beyazıt University	Fadime ÜNEY YÜKSEKTEPE, Ph.D. Prof., Rector, Istanbul Kültür University
Ahmet ŞENGÖNÜL, Ph.D. Prof., Rector, Sivas Cumhuriyet University	Cem ZORLU, Ph.D. Prof., Rector, Necmettin Erbakan University
Nevzat ŞİMŞEK, Ph.D. Prof., Rector, Fatih Sultan Mehmet Vakıf University	Mustafa Latif EMEK, Ph.D. President, Institute of Economic Development and Social Research

Organizing Committee

Sıdıka BAŞÇI, Ph.D. (Chair) Assoc. Prof., Ankara Yıldırım Beyazıt University & Econometric Research Association, Türkiye	Erkan ERDİL, Ph.D. Prof., Middle East Technical University, Türkiye
İbrahim DEMİR, Ph.D. (Co-Chair) Prof., Ankara Yıldırım Beyazıt University, ULİSA-IIRSR, Türkiye	Murat ASLAN, Ph.D. Prof., Ankara Yıldırım Beyazıt University, Türkiye
Mustafa Göktuğ KAYA, Ph.D. (Co-Chair) Assoc. Prof., Ministry of Treasury and Finance, Türkiye	Murad TİRYAKİOĞLU, Ph.D. Assoc. Prof., Afyon Kocatepe University, Türkiye
Fatih Cemil ÖZBUĞDAY, Ph.D. Prof., Ankara Yıldırım Beyazıt University, Türkiye	Ayşenur SAHİNLER, Ph.D. Ankara Yıldırım Beyazıt University, Türkiye

Scientific Committee

Özgür Hakan AYDOĞMUŞ, Ph.D. Prof., Social Sciences University of Ankara, Türkiye	Aykut LENGER, Ph.D. Prof., Ege University, Türkiye
Şükrü APAYDIN, Ph.D. Prof., Nevşehir Hacı Bektaş Veli University, Türkiye	Şaban NAZLIOGLU, Ph.D. Prof., Pamukkale University, Türkiye
Ahmet Faruk AYSAN, Ph.D. Prof., Hamad Bin Khalifa University, Qatar	Nadir ÖCAL, Ph.D. Prof., Çankaya University, Türkiye
Mehmet BALCILAR, Ph.D. Prof., University of New Haven, USA	Özlem ÖNDER, Ph.D. Prof., Ege University, Türkiye
Erkan BAYRAKTAR, Ph.D. Prof., Gulf University of Science & Technology, Kuwait	Mustafa ÖZER, Ph.D. Prof., Anadolu University, Türkiye
Süleyman DEĞİRMEN, Ph.D. Prof., Sivas Cumhuriyet University, Türkiye	Houcine SENOUSSI, Ph.D. Prof., CY Tech, France
Mehmet DEMİRBAĞ, Ph.D. Prof., University of Essex, UK	Ekrem TATOGLU, Ph.D. Prof., Gulf University for Science and Technology, Kuwait
Murat ERTUĞRUL, Ph.D. Prof., Anadolu University, Türkiye	James THEWISSEN, Ph.D. Prof., Université catholique de Louvain, Belgium
Lokman GÜNDÜZ, Ph.D. Prof., Fatih Sultan Mehmet Vakıf University, Türkiye	Nilgün ÇAGLARIRMAK USLU, Ph.D. Prof., Anadolu University, Türkiye
Bülent GÜLOĞLU, Ph.D. Prof., Istanbul Technical University, Türkiye	Asad ZAMAN, Ph.D. Prof., Al-Nafi Online Educational Platform, Pakistan
Christian HAFNER, Ph.D. Prof., Université catholique de Louvain, Belgium	Adam P. BALCERZAK, Ph.D. Assoc. Prof., Universityof Warmia and Mazury, Poland
İhsan IŞIK, Ph.D. Prof., Rowan University, USA	Mustafa KIRCA, Ph.D. Assoc. Prof., Ordu University, Türkiye
Muhsin KAR, Ph.D. Prof., Central Bank of the Republic of Türkiye, Türkiye	Fatma Özgü SERTTAŞ, Ph.D. Assoc. Prof., Ankara Yıldırım Beyazıt University, Türkiye
Mehmet Baha KARAN, Ph.D. Prof., Hacettepe University, Türkiye	Önder ÖZGÜR, Ph.D. Assoc. Prof., Ankara Yıldırım Beyazıt University, Türkiye
Turalay KENC, Ph.D. Prof., The INCEIF University, Malaysia	Saud Ahmed KHAN, Ph.D. Asst. Prof., Pakistan Institute of Development Economics, Pakistan
Yılmaz KILIÇARSLAN, Ph.D. Prof., Anadolu University, Türkiye	Gazi I. KARA, Ph.D. Board of Governors of the Federal Reserve System, USA
Arzdar KIRACI, Ph.D. Prof., Siirt University, Türkiye	Mehmet Fatih ÖZTEK, Ph.D. Ankara Yıldırım Beyazıt University, Türkiye

Conference Secretariat	
Bilge TİRYAKİOĞLU Econometric Research Association (ERA)	Muhammed ORUÇ Res. Asst., Ankara Yıldırım Beyazıt University, ULİSA-IIRSR
Can ALTAY Econometric Research Association (ERA)	Ömer Faruk ACAR Graduate Student, Ankara Yıldırım Beyazıt University
Elmas Bener OTACI Res. Asst., Ankara Yıldırım Beyazıt University, ULİSA-IIRSR	Emre AYIKLAR Student, Ankara Yıldırım Beyazıt University

Supporting Universities

- Fatih Sultan Mehmet Vakıf University, Türkiye
- Necmettin Erbakan University, Türkiye
- İstanbul Kültür University, Türkiye
- Sivas Cumhuriyet University, Türkiye

Supporting Institutions

- INFO Yatırım
- Central Bank of the Republic of Türkiye
- Turkish Airlines
- Tübitak BİLGEM
- Tübitak ULAKBIM
- ODTÜ Mezunlar Birliği Vakfı
- Ankara Düşünce ve Araştırma Merkezi
- Institute of Economic Development and Social Research
- Management and Economics Club, Ankara Yıldırım Beyazıt University, Türkiye
- Artificial Intelligence and Technology Association (Official Representative of US Based Association for the Advancement of Artificial Intelligence Community)

İÇİNDEKİLER

Predicting Systemic Financial Crises via Ensemble Machine Learning Algorithms and SHAP-Based Interpretability

Onur Akkaya [1]

Abstract

We assess how well ensemble machine learning methods predict systemic financial crises in an early-warning setting. The dataset is an annual panel of 23 countries covering 2000 to 2023. A country-year is coded as a crisis when a banking, currency, or debt crisis occurs, and this binary outcome is modelled on macroeconomic, financial, and external-sector indicators. All predictors enter with a one-year lag so that the exercise reflects genuine forecasting rather than concurrent diagnosis. We compare two ensemble approaches: Random Forest, which uses bootstrap aggregation (bagging), and XGBoost, which uses gradient boosting. Both discriminate crises well out of sample, with areas under the ROC curve of 0.941 and 0.949. The Random Forest produces fewer false alarms, while XGBoost ranks vulnerabilities slightly better. Because such models are difficult to interpret, we apply a SHAP (Shapley Additive exPlanations) analysis that reports both the average influence of each indicator and the drivers behind individual predictions, illustrated with two case studies. Four indicators dominate across every attribution method: the prior crisis state, the long-term bond yield, the exchange rate, and the current account balance. These results echo and extend the classical early-warning literature.

Keywords: Systemic Financial Crises, Early Warning Systems, Ensemble Machine Learning, Random Forest, SHAP.

JEL Codes: C53, G01, F37, C55

Topluluk Makine Öğrenmesi Algoritmaları ve SHAP Tabanlı Yorumlanabilirlik ile Sistemik Finansal Krizlerin Tahmini

Özet

Sistemik finansal krizleri erken uyarı çerçevesinde tahmin etmede topluluk (ensemble) makine öğrenmesi yöntemlerinin ne kadar başarılı olduğunu değerlendiriyoruz. Veri seti, 2000-2023 dönemini kapsayan 23 ülkeye ait yıllık bir panelden oluşmaktadır. Bir ülke-yıl gözlemi, bir bankacılık, döviz ya da borç krizi meydana geldiğinde kriz olarak kodlanmakta ve bu ikili (binary) sonuç, makroekonomik, finansal ve dış sektör göstergeleri üzerinden modellenmektedir. Tüm açıklayıcı değişkenler bir yıl gecikmeyle modele girmektedir; böylece çalışma, eşzamanlı bir teşhisten ziyade gerçek bir öngörü egzersizini yansıtmaktadır. İki ensemble yaklaşımını karşılaştırıyoruz: bootstrap aggregation (bagging) yöntemini kullanan Rastgele Orman (Random Forest) ve gradyan artırma (gradient boosting) yöntemini kullanan XGBoost. Her ikisi de örneklem dışında krizleri iyi ayırt etmekte olup, ROC eğrisi altında kalan alan değerleri sırasıyla 0,941 ve 0,949'dur. Rastgele Orman daha az yanlış alarm üretirken, XGBoost kırılganlıkları biraz daha iyi sıralamaktadır. Bu tür modellerin yorumlanması güç olduğundan, hem her bir göstergenin ortalama etkisini hem de bireysel tahminlerin ardındaki etkenleri raporlayan bir SHAP (Shapley Katkı Açıklamaları / Shapley Additive exPlanations) analizi uyguluyor ve bunu iki vaka çalışmasıyla örneklendiriyoruz. Tüm atıf (attribution) yöntemleri genelinde dört gösterge öne çıkmaktadır: önceki kriz durumu, uzun vadeli tahvil getirisi, döviz kuru ve cari işlemler dengesi. Bu sonuçlar, klasik erken uyarı literatürünü yankılamakta ve genişletmektedir.

Anahtar Kelimeler: Sistemik Finansal Krizler, Erken Uyarı Sistemleri, Topluluk Makine Öğrenmesi, Rastgele Orman, SHAP.

JEL Kodları: C53, G01, F37, C55

[1] Onur Akkya, Assoc.Prof. Dr., Department of International Trade and Logistics, Kilis 7 Aralık University, Kilis, Türkiye email: onurakkaya@kilis.edu.tr, ORCID:0000-0003-2694-9073

1. INTRODUCTION

Financial crises recur far more often than their reputation as exceptional events would suggest. The Asian turmoil of 1997–98, the Argentine default of 2001, the global financial crisis of 2008–09, the European sovereign debt crisis of 2010–13, and the emerging-market currency pressures of 2018–19 share a common feature: severe disruption rarely stays within national borders. Trade linkages, cross-border capital flows, and sudden shifts in investor sentiment carry instability from one economy to others, sometimes with consequences for the global financial system as a whole. Because these episodes are both recurrent and contagious, considerable effort has gone into building early-warning systems (EWS) that aim to detect financial fragility from observable macroeconomic and financial indicators before a crisis breaks.

The methods used in this literature have gradually shed their most restrictive assumptions. Kaminsky and Reinhart (1999) introduced the signals approach, which flags an indicator whenever it crosses a threshold calibrated to balance missed crises against false alarms. Around the same time, researchers assembled the historical crisis chronologies, notably those of Reinhart and Rogoff (2009) and Laeven and Valencia (2020), that still anchor most empirical work. A later wave of studies turned to discrete-choice estimators, usually logit or probit, to link crisis incidence to a set of predictors. Both the signals approach and parametric discrete-choice models, however, impose a fairly rigid view of how predictors relate to crisis risk and handle interactions among them only with difficulty. These constraints help explain the recent move toward machine learning, which can represent non-linearities, threshold effects, and interactions without specifying them in advance.

This paper adds to that line of work. It treats systemic crisis prediction as a binary classification problem and applies two related families of ensemble tree methods. Random Forest builds on bootstrap aggregation (bagging) and works mainly by reducing predictive variance; XGBoost builds on gradient boosting and works mainly by reducing predictive bias. We do more than compare their out-of-sample accuracy. We also address a standard objection to such models, namely that they are hard to interpret, by applying a SHAP (Shapley Additive exPlanations) analysis that makes both the overall behaviour of each model and its individual predictions legible.

Three questions guide the analysis. The first is whether ensemble tree methods can discriminate reliably out of sample for an event as rare as a systemic crisis, which occurs in roughly one country-year in twelve. The second is whether bagging and boosting differ in ways that matter in practice, and along which dimensions. The third is whether the indicators the models single out are economically sensible and stable across attribution methods and algorithms. The evidence below supports an affirmative answer in each case and points to interpretable ensemble methods as a useful complement to the existing early-warning toolkit rather than a black-box replacement for it.

The remainder of the paper proceeds as follows. Section 2 situates the study within the literature on crisis prediction and machine learning. Section 3 describes the dataset, the construction of the target variable, and the lagged-predictor design. Section 4 reports the Random Forest analysis and the XGBoost analysis, while evaluates the two paradigms comparatively. and presents the SHAP interpretability analysis, including two case studies. Section 5 discusses the findings and concludes.

2. LITERATURE REVIEW

2.1. The Early-Warning Tradition

Most accounts trace the modern crisis-prediction literature to the signals approach of Kaminsky and Reinhart (1999). They treated each candidate indicator as a binary alarm that sounds once the variable crosses a threshold chosen to balance missed crises against false alarms. Part of the appeal was transparency, since one could read off each indicator's contribution from its own signalling record. The drawback was just as visible: the method judged indicators one at a time and so missed the interactions

among them that often precede a crisis. Alongside this work, the historical cataloguing of Reinhart and Rogoff (2009) and Laeven and Valencia (2020) supplied the consistent, cross-country crisis chronologies that almost all later studies, including this one, rely on to define the dependent variable.

A second generation of studies placed crisis prediction inside discrete-choice econometrics, linking the probability of a crisis to a vector of macroeconomic and financial fundamentals through logit or probit models. Reviewing the indicators that proved informative during the 2008–09 episode, Frankel and Saravelos (2012) found the level of foreign-exchange reserves and the real exchange rate among the most reliable, a result that lines up with the variable-importance rankings reported below. These models delivered interpretable coefficients but remained tied to a linear-in-parameters structure and to the researcher's prior choices about functional form and interaction terms.

2.2. Machine Learning Approaches to Crisis Prediction

The shift toward machine learning reflects a sense that the link between fundamentals and crisis risk is probably non-linear and full of interactions. Tree-based ensembles have been central to this shift. Bagging methods, of which Random Forest (Breiman, 2001) is the standard example, grow many decorrelated trees on bootstrap resamples and average their predictions, which lowers variance and makes overfitting less likely. Boosting methods, set out within the gradient-boosting framework of Friedman (2001) and implemented at scale by Chen and Guestrin (2016) in XGBoost, instead build trees one after another, each correcting the errors of those before it, and so work on bias. A growing body of applied work finds that these ensembles beat logit benchmarks in out-of-sample crisis classification, especially when the predictor set is large and the relationships are non-linear.

The main reservation about these methods concerns interpretability. A logit coefficient has a direct economic reading; the prediction of a 200-tree ensemble does not, and this opacity has slowed their adoption in policy settings, where the reason for a warning has to be stated clearly. Unified attribution frameworks, above all the SHAP method of Lundberg and Lee (2017), have closed much of this gap. Building on the Shapley value from cooperative game theory (Shapley, 1953), SHAP divides a prediction among its features in a way that satisfies sensible axioms of consistency and local accuracy, and it can be computed exactly and quickly for tree ensembles. By providing both global importance rankings and observation-level explanations, SHAP lets the predictive gains of ensemble methods coexist with the interpretability an operational early-warning system requires. This paper sits at that intersection, pairing a careful comparison of bagging and boosting with a SHAP-based reading of both.

3. DATA AND RESEARCH DESIGN

3.1. The Dataset

The analysis uses an annual panel of 23 countries observed from 2000 to 2023, a total of 552 country-year observations. The sample is deliberately mixed. Countries are sorted into five groups according to the role they played in past crises: crisis epicentres such as the United States, Greece, and Argentina; advanced economies that acted as contagion channels, such as Germany, France, the United Kingdom, and Japan; the Fragile Five (Türkiye, Brazil, India, and South Africa); the Fragile Eight, which includes Mexico and Russia; and economies that proved comparatively resilient, such as China, Australia, Canada, Switzerland, and Norway. This grouping enters the analysis as a predictor in its own right and, as Section 5 shows, carries real predictive content for the boosting model.

The explanatory variables fall into four groups, summarised in Table 1. Macroeconomic indicators come from the IMF World Economic Outlook and the World Bank World Development Indicators. Financial-market indicators come from Yahoo Finance, Bloomberg, the Federal Reserve Economic Data service, and the World Bank Global Financial Development Database. External-sector indicators come from UN Comtrade and UNCTAD. The crisis indicators themselves, taken from the chronologies of Laeven and

Valencia (2020) and Reinhart and Rogoff (2009), are not used as predictors; they supply the raw material for the dependent variable instead.

Table 1. Explanatory variable categories and their contents.

Variable category	Contents
Macroeconomic (8)	GDP, GDP growth, CPI inflation, unemployment, current account balance/GDP, public debt/GDP, fiscal balance/GDP, foreign exchange reserves
Financial (4)	Stock market return, year-end exchange rate, 10-year government bond yield, banking-sector Z-score
External trade (6)	Exports, imports, trade balance, trade openness ((X+M)/GDP), FDI inflows, FDI outflows
Crisis indicators (source)	Banking, currency, debt, global-spillover, and Covid crisis flags; the source from which the dependent variable is constructed
Structural (static)	Country group (five categories) and MSCI classification (developed / emerging)

3.2. The Dependent Variable

The outcome is a binary indicator equal to one in any country-year with at least one banking, currency, or debt crisis, and zero otherwise. This composite definition follows the systemic-crisis conventions of Laeven and Valencia (2020) and Reinhart and Rogoff (2009) and is meant to capture the three main channels of systemic instability: trouble in the banking sector, disorder in the foreign-exchange market, and the loss of public-debt sustainability. After the lagged design described below is applied, 43 of the 529 usable observations are crisis years, an incidence of about 8.1 per cent. This kind of class imbalance is inherent to the problem, since systemic crises are infrequent by their nature, and it shapes both the estimation strategy and the choice of evaluation criteria, as Section 3.4 explains.

3.3. The Lagged-Predictor Design

The most important design choice concerns the timing of predictors relative to the outcome. Every explanatory variable enters with a one-year lag, so the crisis status of a given year is predicted only from information available at the end of the previous year. The reasoning is methodological. Admitting contemporaneous values, such as a collapsing equity market or a deteriorating banking Z-score observed during the crisis year itself, would turn the exercise into concurrent diagnosis rather than prediction, and the resulting performance would be inflated by look-ahead bias. The lagged design imposes the discipline of a genuine early-warning system: its outputs correspond to inferences a policymaker could have drawn in real time. Because the first year of each country's series has no lagged value, the modelling sample falls to 529 observations over 2001–2023. The final predictor set has 24 features: eighteen lagged macroeconomic, financial, and trade variables, a lagged indicator of the previous year's crisis status, and indicator variables for country group and MSCI classification.

3.4. Estimation Strategy and Evaluation Criteria

The sample is split, with stratification that preserves the crisis incidence, into an estimation subsample of three-quarters of the observations (396 observations, 32 of them crisis years) and a hold-out subsample of the remaining quarter (133 observations, 11 crisis years). For each algorithm the tuning parameters are chosen on the estimation subsample by a randomised search over a grid, with the area under the ROC curve as the selection criterion and five-fold stratified cross-validation guarding against overfitting to any single partition. The class imbalance is handled inside the estimation procedure: the Random Forest weights observations in inverse proportion to class frequency, and the boosting model

scales the gradient contribution of the minority class by the ratio of non-crisis to crisis observations, roughly eleven to one.

Performance is judged on six measures: accuracy, precision, recall, the F1 score, the area under the ROC curve, and the area under the precision-recall curve. The weight given to the latter measures is deliberate. Under heavy class imbalance, accuracy is a poor guide, since a rule that never predicts a crisis would already reach about 92 per cent accuracy. The area under the ROC curve, the area under the precision-recall curve, and recall speak more directly to the model's ability to tell crises apart from quiet years. To limit the sampling variability that the small number of crisis observations would otherwise introduce, we report performance on the single hold-out partition together with five-fold cross-validated performance over the full sample.

4. METHODOLOGY and ANALYSIS RESULT

4.1. The Random Forest Model

Random Forest (Breiman, 2001) is a bagging method. It grows a large collection of decision trees, each fitted to its own bootstrap resample of the estimation data, and at every node it considers only a randomly chosen subset of the available features. The prediction is the average of the trees' class probabilities. Two features of this design are worth noting. Averaging over many decorrelated trees sharply reduces the high variance of a single tree and makes the predictor naturally resistant to overfitting. And because each tree leaves out roughly a third of the observations from its resample, the predictions on these out-of-bag observations give an internal, nearly unbiased estimate of generalisation performance without the need for a separate validation sample.

The randomised search favoured an ensemble of 200 trees with a maximum depth of twelve, at least six observations per terminal node, at least eight observations required to attempt a split, and half of the features evaluated at each split. This configuration reached a mean area under the ROC curve of 0.870 in cross-validation, and it is the one carried forward to the hold-out evaluation.

Predictive Performance

Table 2 reports the tuned Random Forest on the hold-out sample, and Table 3 the cross-validated figures. On the hold-out sample the model reaches an accuracy of 0.970, an area under the ROC curve of 0.941, and an area under the precision-recall curve of 0.830. The out-of-bag score (0.932) is close to both the hold-out accuracy (0.970) and the area under the ROC curve (0.941), which is reassuring: the model has not simply memorised the estimation data but generalises dependably to observations it has not seen.

Table 2. Random Forest performance on the hold-out sample.

Metric	Value
Accuracy	0.9699
Precision	0.8889
Recall	0.7273
F1 score	0.8000
ROC-AUC	0.9411
PR-AUC	0.8296
Out-of-bag score	**0.9318**

Table 3. Random Forest five-fold cross-validated performance (full sample, n = 529).

Metric (five-fold CV)	Mean ± std. dev.
ROC-AUC	0.8803 ± 0.0573
F1 score	0.5870 ± 0.1422
Recall	0.5611 ± 0.1432
Precision	0.6429 ± 0.1869
Accuracy	0.9339 ± 0.0285

The confusion matrix in Table 4 shows the character of the model's errors. Of the eleven genuine crises in the hold-out sample, eight are caught and three are missed, and only one of the 122 quiet observations is wrongly flagged. The resulting precision of 0.889 means that a crisis warning from the model is usually justified. That property matters in policy contexts where a false alarm can itself be costly, whether by triggering capital controls or by undermining confidence.

Table 4. Random Forest confusion matrix (hold-out sample, n = 133).

	Predicted: no crisis	Predicted: crisis
Actual: no crisis	121	1
Actual: crisis	3	8

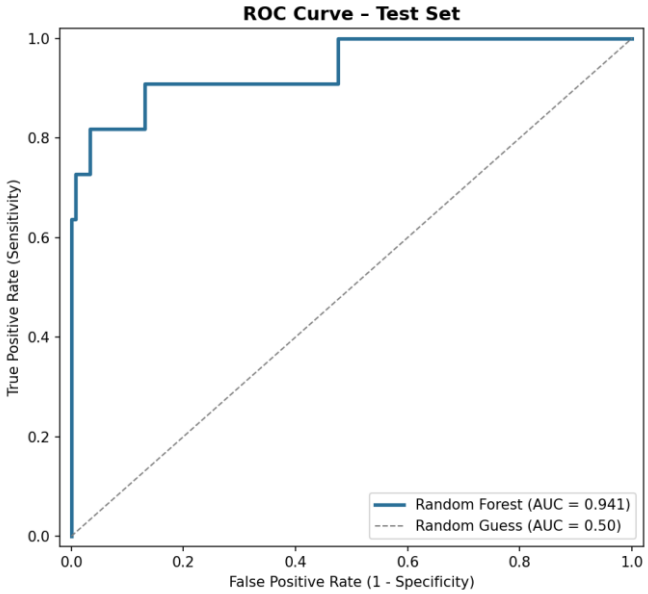


Figure 1. Random Forest ROC curve (hold-out sample, AUC = 0.941).

Variable Importance

Figure 2 shows the variable-importance ranking based on the mean reduction in Gini impurity. The leading predictor is the previous year's crisis status, at 0.175, which reflects how financial crises cluster over several years and tend to recur. Next come the exchange rate (0.134), the ten-year bond yield (0.120), and the current account balance relative to GDP (0.085). Together these four account for about half of the total importance, and they match the indicators stressed in the classical early-warning literature (Kaminsky and Reinhart, 1999; Frankel and Saravelos, 2012). Exchange-rate pressure and

rising long-term yields register the market's pricing of currency and sovereign risk, while the current account balance captures vulnerability on the external-financing side.

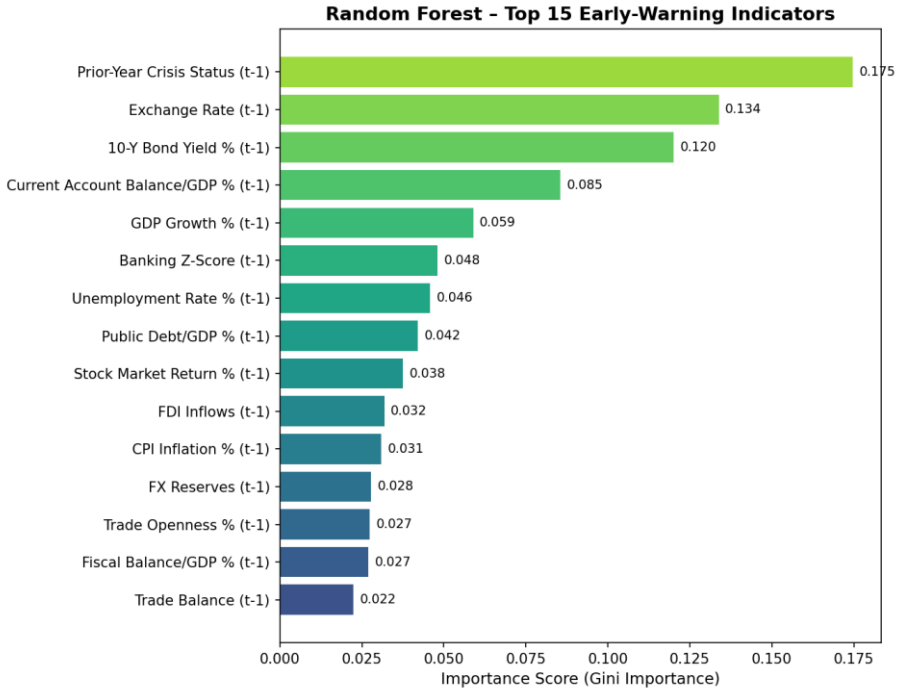


Figure 2. Random Forest variable importance (top fifteen predictors, Gini importance).

4.2. The XGboost Model

XGBoost (Chen and Guestrin, 2016) is a scalable version of the gradient-boosting framework of Friedman (2001). Where Random Forest builds its trees in parallel and independently, boosting builds them in sequence: each new tree is fitted to the residual errors, that is, the gradient of the loss function, left by the trees assembled so far. By working on bias directly, this approach often sharpens discrimination, but it is also more prone to overfitting, so the regularisation parameters have to be tuned with care.

To handle the imbalance between crisis and quiet observations, the minority class's contribution to the gradient is scaled by the ratio of non-crisis to crisis years, roughly eleven to one. The randomised search favoured an ensemble of 200 trees with a maximum depth of five, considerably shallower than the Random Forest to keep overfitting in check, a learning rate of 0.03, a row-subsampling rate of 0.8, a column-subsampling rate of 0.6, a minimum child weight of five, and an L2 regularisation coefficient of two. This configuration reached a mean area under the ROC curve of 0.873 in cross-validation.

Predictive Performance

On the hold-out sample the boosting model reaches an area under the ROC curve of 0.949, just above the 0.941 of the Random Forest, and the edge holds in cross-validation (0.891 against 0.880). Its precision, though, drops to 0.667 from the Random Forest's 0.889, which corresponds to four false alarms in the hold-out sample rather than one (Table 6). Both models recover the same eight of eleven genuine crises, so the gap lies not in sensitivity but in the false-alarm rate.

Table 5. XGBoost performance on the hold-out sample.

Metric	Value
Accuracy	0.9474
Precision	0.6667
Recall	0.7273
F1 score	0.6957
ROC-AUC	0.9493
PR-AUC	0.8184

Table 6. XGBoost five-fold cross-validated performance (full sample, n = 529).

Metric (five-fold CV)	Mean ± std. dev.
ROC-AUC	0.8914 ± 0.0569
F1 score	0.5908 ± 0.1629
Recall	0.5833 ± 0.1648
Precision	0.6076 ± 0.1766
Accuracy	0.9339 ± 0.0286

Table 7. XGBoost confusion matrix (hold-out sample, n = 133)

	Predicted: no crisis	Predicted: crisis
Actual: no crisis	118	4
Actual: crisis	3	8

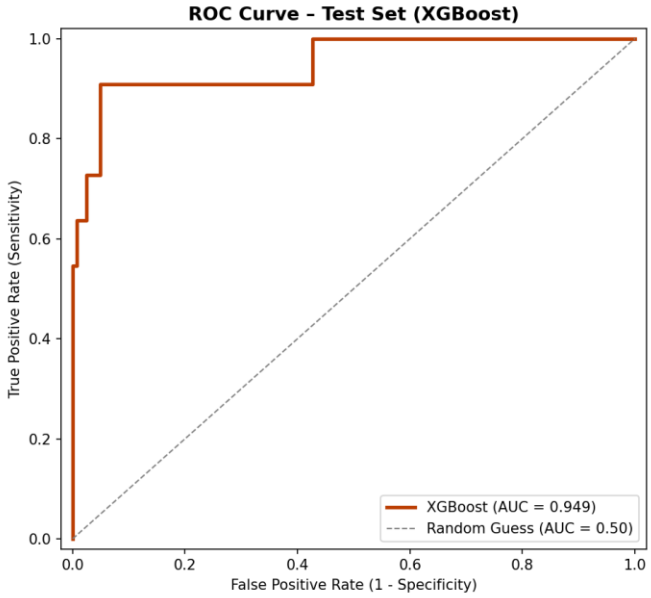


Figure 3. XGBoost ROC curve (hold-out sample, AUC = 0.949).

Variable Importance

The gain-based importance ranking of the boosting model (Figure 4) overlaps a good deal with the Random Forest, but differs in one telling way: the indicator for the Fragile Eight group ranks second, just behind the previous year's crisis status, at 0.092. This suggests that the sequential, error-correcting logic of boosting can pin residual risk, the part left unexplained by the individual macroeconomic

indicators, on a country's structural fragility grouping. The bagging architecture surfaces this kind of cross-country heterogeneity less prominently.

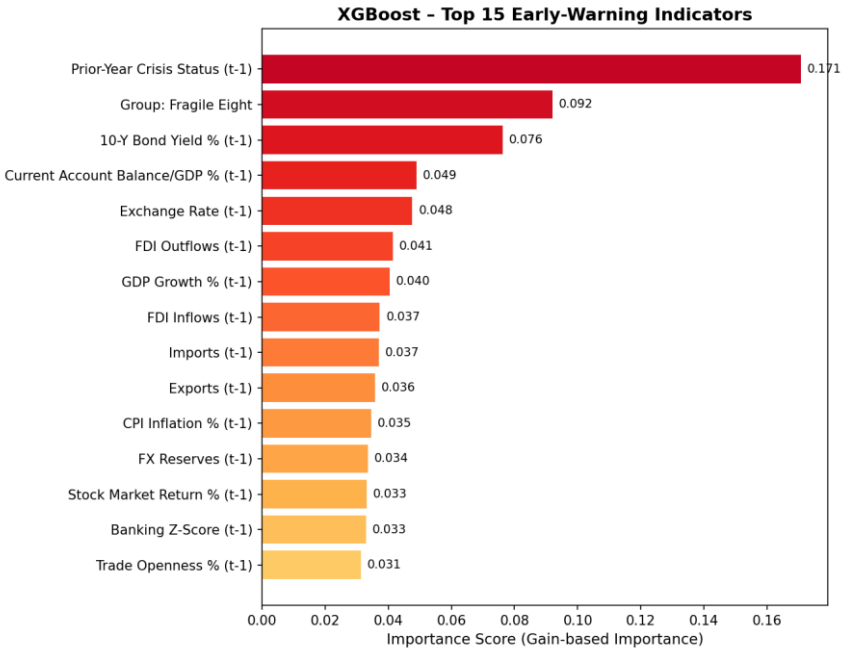


Figure 4. XGBoost variable importance (top fifteen predictors, gain importance).

4.3. Bagging Versus Boosting: A Comparative Assessment

Because both models were estimated on the same sample, with the same definition of the dependent variable and the same split into estimation and hold-out subsamples, their results can be compared directly. Table 8 puts the hold-out figures side by side, and Figures 5 and 6 show the two ROC curves and a comparison of the six performance measures.

Table 8. Hold-out performance: Random Forest (bagging) versus XGBoost (boosting).

Metric	Random Forest	XGBoost	Difference
Accuracy	0.9699	0.9474	−0.0225
Precision	0.8889	0.6667	−0.2222
Recall	0.7273	0.7273	0.0000
F1 score	0.8000	0.6957	−0.1043
ROC-AUC	0.9411	0.9493	+0.0082
PR-AUC	0.8296	0.8184	−0.0112

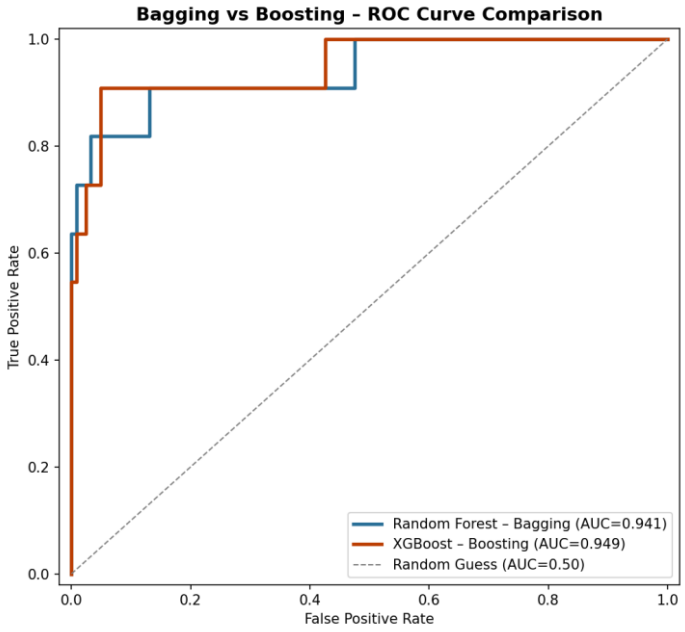


Figure 5. Comparison of ROC curves, bagging versus boosting.

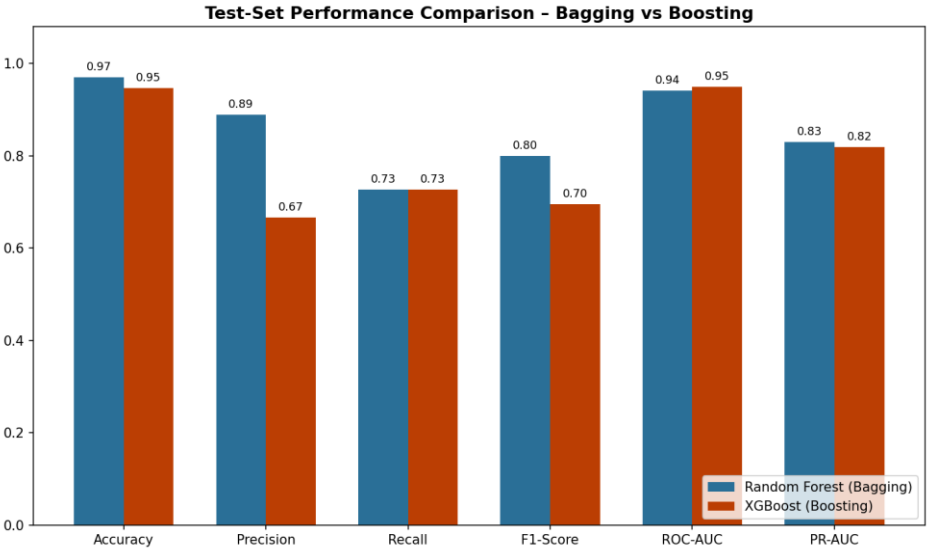


Figure 6. Comparison of hold-out performance measures across methods.

The two approaches discriminate about equally well overall on this sample, with areas under the ROC curve of roughly 0.94 to 0.95, but their errors differ in a systematic way. The Random Forest issues fewer false alarms and so classifies more conservatively, while the boosting model ranks vulnerabilities slightly better. This contrast follows from the distinctions in Table 9. The defining aim of bagging is to reduce variance; that of boosting is to reduce bias. With a modest sample and sharp class imbalance, the variance-reducing approach gains in precision, and the bias-reducing approach gains a little in discrimination.

Table 9. Conceptual contrasts between bagging and boosting and their manifestation in this study.

Dimension	Random Forest (bagging)	XGBoost (boosting)
Tree construction	Parallel, independent (bootstrap)	Sequential, error-correcting
Statistical objective	Variance reduction	Bias reduction
Internal validation	Out-of-bag score available	Requires separate CV / early stopping
Over-fitting tendency	Lower; stable in tree count	Higher; needs careful regularisation

Dimension	Random Forest (bagging)	XGBoost (boosting)
Outcome here	Higher precision, fewer false alarms	Marginally higher-ranking power

The point is not that one method beats the other, but that the choice depends on the relative cost of the two kinds of error in the intended use. Where a false alarm is expensive, as in an official warning system whose statements set off intervention, the higher precision of the bagging model recommends it. Where the goal is instead to rank economies by relative vulnerability, as in surveillance or triage, the slightly better discrimination of the boosting model is the more relevant strength.

4.4. Interpretability: A SHAP-Based Analysis

Gini and gain importance measures summarise the average influence of a variable on the model, but they say nothing about the direction of that influence or about how it plays out in a particular case. To fill this gap, the analysis uses the SHAP framework of Lundberg and Lee (2017), which rests on the Shapley value from cooperative game theory (Shapley, 1953). For each observation, SHAP measures how far every feature moves the prediction away from a baseline expectation, and these contributions sum exactly to the realised prediction. For tree ensembles the values can be computed exactly and efficiently, and the framework delivers both a global account of variable influence and a breakdown of individual predictions.

Two cautions apply to the figures that follow. Because the Random Forest weights the classes during estimation, its SHAP baseline reflects the reweighted internal class distribution rather than the raw crisis incidence; the additive decomposition still holds exactly, and the direction and ranking of effects stay valid. The SHAP values for the boosting model, in turn, are on the log-odds scale rather than the probability scale, so the magnitudes are not directly comparable across the two models even though the qualitative reading is.

Global Effects

Figures 7 and 8 show the SHAP summary plots for the two models. In each, a point is a single country-year, its colour encodes the value of the feature, and its horizontal position encodes the sign and size of the feature's contribution to the prediction. The patterns fit economic intuition: high long-term bond yields, which register the market's pricing of risk, move the prediction toward crisis, while a stronger current account position moves it away from crisis.

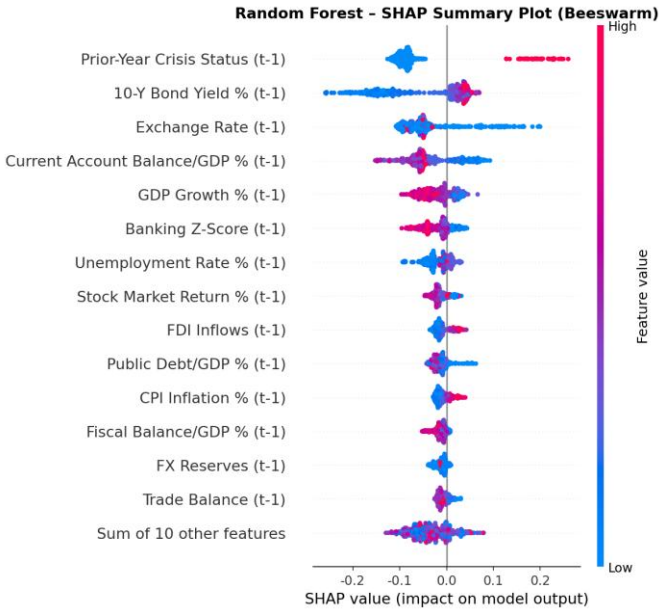


Figure 7. Random Forest SHAP summary plot.

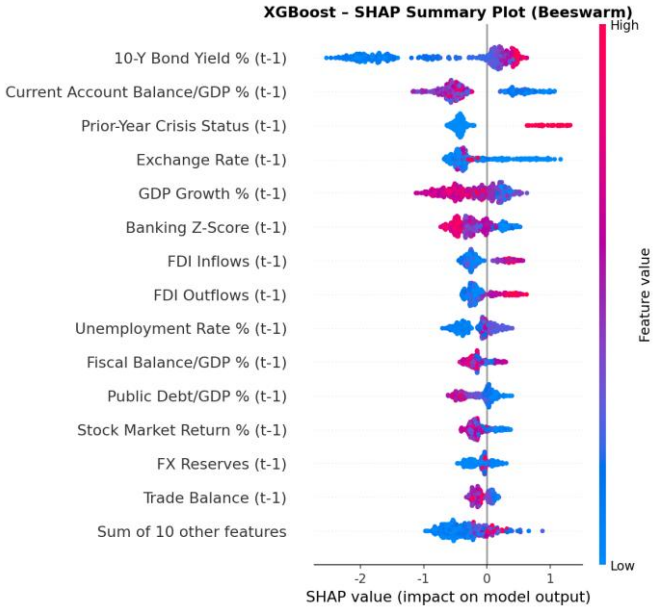


Figure 8. XGBoost SHAP summary plot.

The same four variables, namely the prior crisis state, the ten-year bond yield, the exchange rate, and the current account balance, lead the SHAP ranking of both models (Tables 10 and 11). That this set recurs across two different importance criteria, Gini and gain on one side and SHAP on the other, and across two very different algorithms, is strong evidence that the leading indicators are not a by-product of any one method or sampling draw but reflect a real and stable structure in the data.

Table 10. Random Forest: leading variables by mean absolute SHAP value.

Rank	Variable	Mean |SHAP|
1	Prior-year crisis status	0.0991
2	10-year bond yield	0.0729
3	Exchange rate	0.0670
4	Current account balance/GDP	0.0550
5	GDP growth	0.0301

Table 11. XGBoost: leading variables by mean absolute SHAP value.

Rank	Variable	Mean |SHAP|
1	10-year bond yield	0.8138
2	Current account balance/GDP	0.5573
3	Prior-year crisis status	0.4899
4	Exchange rate	0.4278
5	GDP growth	0.3517

5. DISCUSSION and CONCLUSION

Three themes run through the results. The first is predictive adequacy. Both ensembles discriminate crises well out of sample, with areas under the ROC curve around 0.94 to 0.95, for an event that occurs in only about eight per cent of country-years. That tree-based methods, which can capture the non-

linearities, threshold effects, and interactions that linear models miss, perform this well suggests they are a worthwhile addition to the early-warning toolkit. A second theme is the robustness of the indicators. The conventional importance measures and the SHAP attributions, computed separately across two different algorithms, point to the same four leading indicators: crisis recurrence, the long-term bond yield, exchange-rate pressure, and the current account balance. This agreement makes it hard to dismiss the findings as an artefact of method or sample. It also lines up with the indicators stressed in the classical literature (Kaminsky and Reinhart, 1999; Frankel and Saravelos, 2012), which suggests that machine learning here does not overturn the received view so much as confirm and extend it. The third theme is the contrast between the two approaches, which is less a question of which is better than a trade-off that depends on context. The higher precision of the bagging model suits cases where false alarms are costly, while the slightly better discrimination of the boosting model suits cases where the relative ranking of vulnerabilities is what matters. There is also a practical point about transparency. Ensemble methods are often dismissed as opaque, but the Russian episode shows how SHAP works against that view: seeing why a model fails reveals which kinds of information, here exogenous and geopolitical shocks, it leaves out, and that gives both the analyst and the policymaker a concrete agenda for improvement.

This paper applied and compared two ensemble approaches, Random Forest based on bagging and XGBoost based on boosting, to the prediction of systemic financial crises across 23 economies from 2000 to 2023. A lagged-predictor design set up a genuine early-warning framework, and within it both models discriminated crises well out of sample, with areas under the ROC curve above 0.94. The bagging model favoured precision and so issued fewer false alarms, while the boosting model ranked vulnerabilities slightly better, a contrast that follows directly from the two methods' emphases on reducing variance and reducing bias. The SHAP analysis added a layer of interpretability that mattered. It showed the direction and size of the indicator effects overall, and through the Greek and Russian cases it exposed the concrete reasons behind the models' hits and misses. Across four separate attribution exercises, the same four indicators stood out: the prior crisis state, the long-term bond yield, exchange-rate pressure, and the current account balance. Their consistency speaks to the robustness of the findings and to their fit with the established early-warning literature. On this evidence, interpretable ensemble methods look like a promising tool for building early-warning systems for systemic financial crises, combining solid predictive performance with enough transparency to meet the needs of policy.

REFERENCES

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Frankel, J. and G. Saravelos (2012). Can Leading Indicators Assess Country Vulnerability? Evidence from the 2008–09 Global Financial Crisis. *Journal of International Economics*, 87(2), 216–231.

Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189–1232.

Kaminsky, G.L. and C.M. Reinhart (1999). The Twin Crises: The Causes of Banking and Balance-of-Payments Problems. *American Economic Review*, 89(3), 473–500.

Laeven, L. and F. Valencia (2020). Systemic Banking Crises Database II. *IMF Economic Review*, 68, 307–361.

Lundberg, S.M. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Reinhart, C.M. and K.S. Rogoff (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton: Princeton University Press.

Shapley, L.S. (1953). A Value for n-Person Games. In *Contributions to the Theory of Games*, Vol. II, ed. H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press, 307–3.

A Dual-Model System for Vehicle Service Retention: Predicting Service Timing and Churn Probability

Selçuk Bayracı [1] and Garen Bozoğlanoğlu [2]

Abstract

Vehicle service centers face significant challenges in customer retention and operational planning. We present a dual-model system that addresses both challenges simultaneously by predicting (1) the expected next service arrival time and (2) the probability of service churn. Our system employs two LightGBM models trained on historical service records, vehicle characteristics, and customer behavior data. The first model forecasts when a customer is likely to return for their next service, enabling proactive scheduling and resource allocation. The second model estimates churn probability, identifying at-risk customers who may defect to competitors or discontinue service. By combining temporal predictions with retention risk assessment, our system enables targeted, time-sensitive interventions that align operational capacity with customer retention strategies. This integrated approach bridges the gap between maintenance demand forecasting and churn prediction, two domains that have traditionally been addressed separately in the literature.

Keywords: Customer Churn Prediction, Service Timing Forecasting, LightGBM, Automotive After-Sales, Probability Calibration

JEL Codes: C53, C25, M31, L62

Araç Servis Bağlılığı İçin İkili Model Sistemi: Servis Zamanlaması ve Kayıp (Churn) Olasılığının Tahmin Edilmesi

Özet

Araç servis merkezleri, müşteri elde tutma ve operasyonel planlama konularında önemli zorluklarla karşılaşmaktadır. Bu çalışmada, (1) beklenen bir sonraki servis geliş zamanını ve (2) servisi bırakma (churn) olasılığını tahmin ederek her iki zorluğu aynı anda ele alan çift modelli bir sistem sunuyoruz. Sistemimiz; geçmiş servis kayıtları, araç özellikleri ve müşteri davranış verileri üzerinde eğitilmiş iki LightGBM modeli kullanmaktadır. İlk model, bir müşterinin bir sonraki servisi için ne zaman dönme ihtimalinin yüksek olduğunu öngörerek proaktif planlama ve kaynak tahsisine olanak tanır. İkinci model ise servisi bırakma olasılığını tahmin ederek, rakiplere geçebilecek veya hizmet almayı durdurabilecek risk altındaki müşterileri tespit eder. Sistemimiz, zamansal tahminleri elde tutma riski değerlendirmesiyle birleştirerek, operasyonel kapasiteyi müşteri elde tutma stratejileriyle uyumlu hale getiren, hedefe yönelik ve zamana duyarlı müdahalelere olanak sağlar. Bu entegre yaklaşım, literatürde geleneksel olarak ayrı ayrı ele alınan iki alan olan bakım talebi tahmini ile müşteri kaybı (churn) tahmini arasındaki boşluğu doldurmaktadır.

Anahtar Kelimeler: Kayıp Tahmini, Müşteri Bağlılığı, Talep Tahmini, Öngörücü Modelleme, LightGBM, Araç Servis Merkezleri.

JEL Kodları: C53, M31, M11.

1. INTRODUCTION

Customer retention in the automotive service industry directly impacts revenue stability and operational efficiency. Service centers must balance two critical objectives: accurately forecasting service demand to optimize resource allocation, and identifying at-risk customers to implement timely retention interventions. Traditional approaches address these objectives independently, treating service timing prediction and churn analysis as separate problems.

[1] Selçuk Bayracı, R&D Center, Borusan Otomotiv, İstanbul, Türkiye
e-mail: selcuk.bayraci@borusanotomotiv.com, ORCID: 0000-0003-4831-4802
[2] Garen Bozoğlanoğlu, Data Analytics & AI Dept., Borusan Otomotiv, Istanbul, Türkiye
e-mail: garen.bozoglanoglu@borusanotomotiv.com. ORCID: 0009-0001-4977-6972

We propose a dual-model system that integrates both objectives into a unified retention framework. Our system employs two LightGBM models: one predicting the expected time until the next service visit, and another estimating the probability that a customer will churn before that visit occurs. This combination enables service centers to prioritize retention efforts based on both timing urgency and churn risk, directing resources toward customers who are both likely to churn and approaching their expected service window.

The automotive service retention domain presents unique challenges compared to traditional churn prediction contexts. Service visits are inherently temporal and mileage-dependent, with customers returning at irregular intervals driven by vehicle usage patterns, maintenance schedules, and repair needs. Furthermore, the relationship between service history, vehicle characteristics, and future behavior creates complex dependencies that require sophisticated modeling approaches.

Our contributions are threefold: (1) we demonstrate the application of LightGBM for both service timing and churn prediction in an integrated system, (2) we show how combining these predictions enables more effective retention strategies than either model alone, and (3) we provide a practical framework that service centers can deploy to improve customer retention and operational planning.

2. PROBLEM STATEMENT

2.1. The Economic Centrality of Automotive After-Sales Services

The automotive after-sales service market has emerged as one of the most strategically significant segments of the global automotive value chain. The market was valued at approximately USD 1.0 trillion in 2023 and is projected to grow at a compound annual rate of 5% to reach USD 1.4 trillion by 2030 (Verified Market Research, 2024). Of particular note is the disproportionate contribution of after-sales operations to dealer profitability: although after-sales activities represent only approximately 10% of total dealership revenue, they account for more than 40% of gross profit for large full-range dealerships (Spyne, 2024). An early industry benchmark established by Accenture documented that General Motors generated greater profits from after-sales revenues in a single year than from USD 150 billion in vehicle sales, underscoring the structural importance of the service stream relative to new-car transactions (ProCount West, 2018).

This margin asymmetry is grounded in the economics of customer lifetime value (CLV). In the automotive sector, CLV extends well beyond the initial vehicle purchase and is primarily realised through recurring service events across the ownership lifecycle. Reichheld and Sasser (1990) demonstrated in a seminal study that even a modest reduction in customer defection rates can produce disproportionately large gains in long-term profitability, a result that has since been corroborated in automotive-specific contexts (Kumar and Reinartz, 2018). Publicis Sapient (2023) observe that customers tend to remain loyal for the first three years of ownership due to warranty lock-in, but that loyalty decays sharply in the post-warranty period in the absence of proactive relationship management, revealing a critical retention window that authorised service centres must actively defend.

Compounding this challenge, brand loyalty in the automotive sector has weakened considerably in recent years, declining from 54% in 2019 to 49.2% in 2022 — the lowest level recorded since 2014 (IHS Markit, 2022). The proliferation of independent quick-lube chains, multi-brand service networks, and digitally-enabled mobility services has created a competitive landscape in which customers face low switching costs and abundant alternatives. Only approximately one-third of consumers report a preference for returning to an authorised dealership service centre (Cox Automotive, 2021), and this proportion falls further when service inconvenience, pricing opacity, or poor communication are experienced. Against this backdrop, the ability to identify at-risk customers before they defect — and to intervene at the precise moment when a service visit is due — constitutes a core competitive capability for authorised vehicle service networks.

2.2. The Non-Contractual Nature of Automotive Service Churn

Customer churn in the automotive after-sales context is structurally distinct from churn in contractual settings such as telecommunications or subscription services. In contractual environments, a churn event is directly observable: a customer explicitly terminates a contract, and this event is recorded in the database. In the automotive service domain, no such contract governs the service relationship. A customer who decides to take their vehicle to a competitor simply stops returning, without any notification. This is the defining characteristic of the non-contractual churn problem, first formalised by Fader and Hardie (2007) in the Pareto/NBD modelling tradition.

In non-contractual settings, the churn event must be constructed rather than observed. A common and practically grounded approach is to define a customer as churned if no return visit is recorded within a pre-specified inactivity threshold τ (van Meeuwen, 2021). The appropriateness of this threshold is domain-specific: in automotive service, the manufacturer-recommended service interval typically spans 12 to 24 months, making a threshold of $\tau = 15$ months a principled choice. Formally, for vehicle v with its j-th service at time t_j, the churn indicator is defined as:

$$C(v, j) = \mathbb{1}[\, t_{j+1} - t_j > \tau \;\; \vee \;\; t_{j+1} = \emptyset \,] \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function and the second condition captures vehicles that have not yet returned at the time of scoring. This construction transforms an inherently censored observation problem into a tractable binary classification task.

A further structural challenge is that the population of active vehicles at any given time contains both fully-resolved historical service sequences and right-censored sequences for which the next service date is unknown. This censoring structure motivates the strict temporal separation between training and scoring populations adopted in our methodology, following the forward-in-time validation principle advocated by Neslin et al. (2006).

2.3. Limitations of Existing Approaches and the Research Gap

The academic literature on data-driven customer churn prediction is extensive, with comprehensive surveys documenting applications across telecommunications, banking, insurance, e-commerce, and gaming (Ahn et al., 2020; Ullah et al., 2025; Shahabikargar et al., 2026). Machine learning approaches — particularly ensemble methods such as random forests, gradient-boosted trees, and, more recently, LightGBM — have consistently outperformed classical logistic regression and survival analysis baselines in discriminative power (Lemmens and Gupta, 2020; Liu et al., 2024).

Within the automotive after-sales domain specifically, the literature remains comparatively sparse. Wang et al. (2020) is the most directly relevant prior work, addressing customer absence prediction for automobile 4S shops using a recurrent neural network variant designed to exploit the lifecycle structure of longitudinal service sequences. However, Wang et al. (2020) address only the binary question of whether a customer will be absent, and do not estimate when an active vehicle is expected to return.

More broadly, the extant literature on non-contractual churn prediction has addressed the timing and defection problems separately and in isolation. Survival analysis methods model the time-to-event distribution and are well-suited to the timing problem, but their standard formulations assume a single terminal event and do not naturally accommodate recurrent service visits (Larivière and Van den Poel, 2004; Geiler et al., 2022). Conversely, binary churn classifiers model defection probability as a static outcome and discard the temporal structure of the inter-visit distribution. The dual challenge — predicting when a vehicle will return and whether it will return at all — thus constitutes an unaddressed joint inference problem in the automotive service literature.

2.4. Formal Problem Statement

Let $V = \{v_1, \ldots, v_N\}$ be the set of vehicles with at least one periodic maintenance service record. For each vehicle $v \in V$, let $S(v) = \{(t_1, x_1), \ldots, (t_{n_v}, x_{n_v})\}$ denote the ordered sequence of periodic maintenance service events, where t_k is the service timestamp in days and x_k is a feature vector. Let $A \subseteq V$ denote the set of active vehicles at the time of scoring.

The system is required to solve two prediction problems for each vehicle $v \in A$ simultaneously:

> *Problem 1 (Service Timing).* Estimate the conditional distribution $P(Y \mid x_{n_v})$ over the inter-service duration $Y = (t_{n_v+1} - t_{n_v}) / 30$, discretised to $K = 15$ ordinal classes corresponding to months 1 through 15. From this distribution, derive a point prediction $k(x)$ *and an expected arrival date* $EAD(v) = t_{n_v} + k(x) \times 30$ days.
>
> *Problem 2 (Churn Probability).* Estimate the conditional probability $P(C = 1 \mid x_{n_v})$ that vehicle v will not return within τ months of its most recent service. From this probability, derive a calibrated churn score $ChurnRate(v) \in [0,1]$ and its complement $RetentionRate(v) = 1 - ChurnRate(v)$.

These two problems are coupled: the expected arrival date from Problem 1 enters the feature space of Problem 2 as a covariate, and the joint outputs are consumed together by the retention management system.

2.5. Why a Dedicated Dual-Model Architecture Is Necessary

One might ask whether an existing unified framework — such as a survival model or a BTYD specification — could address both sub-problems jointly. We argue that such approaches are inadequate for the present setting on three grounds.

Rich covariate structure. BTYD models are parsimonious by design, relying on frequency, recency, and age as sufficient statistics. In the automotive service context, however, vehicle-level features (engine displacement, horsepower, model year), service content features, and financial covariates carry substantial predictive information beyond RFM. Machine learning methods with gradient-boosted trees have been shown to outperform statistical baselines when rich, heterogeneous feature sets are available (Chen and Guestrin, 2016).

Separation of timing and defection objectives. The distributions governing inter-service timing and defection probability are governed by different underlying processes. The timing distribution is approximately unimodal, making an ordinal classification treatment appropriate, while the defection distribution is structurally sparse, making a calibrated binary classifier more appropriate.

Operational asymmetry. The outputs of the two sub-problems are consumed by different downstream business processes: expected arrival dates feed into workshop scheduling and parts procurement systems, while churn scores feed into CRM campaign targeting systems. A modular dual-model architecture is therefore preferable to a monolithic joint model.

3. METHODOLOGY

This section formalises the probabilistic framework underlying our dual-model architecture. The pipeline is structured as a directed acyclic graph of computational stages (preprocess → km_pred → model → metrics → log → db_writer), each operating on immutable intermediate datasets serialised in Apache Parquet format.

3.1. Problem Formulation and Notation

Let $V = \{v_1, v_2, \ldots, v_N\}$ denote the set of vehicles. For each vehicle $v \in V$, let the ordered sequence of periodic maintenance service events be $S(v) = \{(t_1, x_1), (t_2, x_2), \ldots, (t_{n_v}, x_{n_v})\}$, where t_k is the service timestamp in days and x_k is the corresponding feature vector. Let $T_k = t_{k+1} - t_k$ denote the inter-service interval in days.

We define the target inter-service duration in months as:

$$Y_k = (t_{k+1} - t_k) / 30, \quad Y_k \in \mathbb{R}_+ \tag{2}$$

A vehicle v is said to have churned after its j-th service if:

$$C(v, j) = \mathbb{1}[\, t_{j+1} > t_j + \tau \,] \tag{3}$$

where $\tau = 15$ months is the churn threshold. If no subsequent service is recorded, $C(v, j) = 1$ by definition.

3.2. Data Preprocessing and Feature Engineering

3.2.1. Imputation of First-Visit Records

For vehicles making their inaugural service visit (PMServiceCount = 1), the inter-visit mileage accumulation and elapsed time cannot be computed from prior visits. We adopt the following deterministic imputation rules:

$$KmAccSinceLastVisit_1 = Mileage_1 \quad (if\ missing) \tag{4}$$

$$TimeElapsedSinceLastVisit_1 = (t_1 - FR_v) / 30 \quad (if\ missing) \tag{5}$$

3.2.2. Quality Filters

Service instances satisfying any of the following conditions are excluded from model training: (F1) negative mileage accumulation; (F2) zero or minimal elapsed time; (F3) missing engine horsepower or displacement; (F4) services covered by BSI or Borusan Oto fleet programmes. These filters remove measurement errors and structurally different service events.

3.2.3. Feature Space

After filtering, the feature vector x_k encompasses five categories of predictors: (i) vehicle static attributes (engine horsepower, engine capacity, vehicle age since first registration); (ii) usage and temporal covariates (mileage since last visit, elapsed time, periodic maintenance service count); (iii) historical service expenditure over the prior two years; (iv) service content indicators (binary flags for oil, filters, spark plugs, etc.); and (v) warranty status covariates.

3.3. Service Timing Prediction: A Probabilistic Ordinal Classification Framework

3.3.1. Discretisation of the Inter-Service Duration

Although Y_k is continuous in principle, its distribution is strongly concentrated on integer multiples of calendar months due to manufacturer-recommended service intervals. We discretise Y_k into $K = 15$ ordinal classes by rounding to the nearest integer and capping at 15:

$$M\_k = min(\ round(Y\_k),\ 15\ ),\quad M\_k \in \{1,\ 2,\ \ldots,\ 15\} \tag{6}$$

3.3.2. Multiclass LightGBM Classifier

We model the conditional class probability vector as:

$$p(m \mid x) = P(M = m \mid x) = Softmax(f\_m(x)),\quad m \in \{0,\ \ldots,\ 14\} \tag{7}$$

where f_m : $\mathbb{R}$^d → $\mathbb{R}$ is the m-th output of a gradient-boosted decision tree ensemble trained with the multiclass cross-entropy (softmax log-loss) objective:

$$\mathscr{L}\_timing = -\ (1/n)\ \Sigma_i\ \Sigma_m\ \mathbb{1}[\tilde{M}_i = m] \cdot log\ p(m \mid x_i) \tag{8}$$

The ensemble is parameterised with 50 leaves per tree, a learning rate of η = 0.05, gradient-based one-sided sampling (GOSS) with a bagging fraction of 0.3 and bagging frequency of 30, and a feature subsampling ratio of 0.8.

3.3.3. Continuous Ranked Probability Score

Point accuracy metrics are insufficient for evaluating probabilistic forecasts over ordinal targets. We adopt the Continuous Ranked Probability Score (CRPS), which measures the integrated squared deviation between the predicted cumulative distribution function and the empirical CDF:

$$CRPS(\hat{p}_i,\ m_i) = \Sigma_{k=0}{}^{14}\ (\ \hat{F}_i(k) - \mathbb{1}[m_i \le k]\ )^2 \tag{9}$$

3.3.4. Median-Based Point Prediction and ±1-Month Accuracy

A point prediction is derived from the predicted CDF as the class-index median, defined as the first class k* for which the cumulative probability exceeds 0.5:

$$k^*(x) = min\ \{\ k \in \{0,\ \ldots,\ 14\} : \hat{F}(k \mid x) > 0.5\ \} \tag{10}$$

The ±1-month accuracy is then:

$$Acc{\pm}1 = (1/n)\ \Sigma_i\ \mathbb{1}[\ |k^*(x_i) - m_i| \le 1\ ] \tag{11}$$

3.3.5. Expected Arrival Date

The predicted expected arrival date for vehicle v, given its most recent service date t_j, is computed as:

$$EAD(v) = t_j + (\ k^*(x_j) + 1\ ) \times 30\ days \tag{12}$$

3.4. Churn Probability Estimation: Binary Classification with Probability Calibration

3.4.1. Churn Label Construction and Temporal Validity

Let t_ref denote the reference cutoff date, defined as the most recent service date in the dataset minus the churn threshold τ = 15 months. A service instance (v, j) is included in the labelled training corpus only if t_j < t_ref, ensuring that the full 15-month window has elapsed and the churn outcome is fully observed.

3.4.2. Temporal Train / Validation / Test Splits

The labelled corpus is partitioned into three non-overlapping temporal folds. The minimum training date is fixed at t_min = 2016-10-01. Train: { (v,j) : t_min ≤ t_j < t_churn − Δ_val }; Validation: { (v,j) : t_churn − Δ_val ≤ t_j < t_churn }; Test: { (v,j) : t_j ≥ t_churn }, retaining only the most recent service per vehicle. This strictly forward-in-time splitting prevents any future information from contaminating the training sets.

3.4.3. Class Imbalance Correction

Churn events are structurally rare. To avoid the model collapsing to the majority class, the negative set is randomly under-sampled without replacement to yield a balanced training set with |D_train| = $2n_+$. The LightGBM trainer additionally receives the flag is_unbalance = True, which internally adjusts the gradient magnitudes by the inverse class frequency.

3.4.4. Gradient-Boosted Binary Classifier

The churn model minimises the binary log-loss objective:

$$\mathcal{L}\_churn = -(1/n)\,\Sigma_i\,[\,y_i \log \sigma(f_i) + (1-y_i)\log(1-\sigma(f_i))\,] \tag{13}$$

The LightGBM hyperparameters — determined by prior Bayesian optimisation — are: 33 leaves, feature fraction 0.524, bagging fraction 0.5 with bagging frequency 56, and learning rate 0.158.

3.4.5. Platt Scaling for Probability Calibration

Raw LightGBM sigmoid outputs are known to be poorly calibrated when the training distribution is artificially balanced. We apply Platt scaling, a parametric post-hoc calibration method, to transform raw scores into calibrated probabilities. A logistic regression model is fitted on the training data:

$$P\_cal(y=1 \mid \lambda) = \sigma(w_0 + w_1 \cdot \lambda) \tag{14}$$

At inference time, the fitted calibration model yields calibrated churn probabilities: ChurnRate(v) = P_cal(y=1 | λ(v)) ∈ [0, 1].

3.4.6. Decision Threshold and Performance Metrics

For computing classification metrics, a decision threshold θ = 0.60 is applied. The elevated threshold (relative to the canonical 0.5) reflects the asymmetric cost structure: false positives (unnecessary outreach) are less costly than false negatives (undetected at-risk customers who subsequently churn). Reported metrics include recall and precision for the positive class.

3.5. Retention Score and Risk Ranking

The retention score for vehicle v is defined as the complement of the calibrated churn probability: RetentionRate(v) = 1 − ChurnRate(v). Vehicles are ranked in descending order of ChurnRate(v), producing an ordinal risk rank over the set A of active vehicles. This ranking enables the customer retention team to prioritise outreach interventions proportional to churn risk.

3.6. Output Integration

The outputs of the two models are joined on the composite key (VIN, ServiceDate) to produce a unified record per active vehicle. Three disjoint vehicle subsets are identified and reported separately:

already-churned vehicles (whose inactivity threshold has elapsed), BSI/Borusan-covered vehicles (whose service behaviour is externally mandated), and active vehicles (for whom the full dual-model output is computed).

4. EMPIRICAL DATA ANALYSIS AND RESULTS

4.1. Dataset and Evaluation Cohort

The empirical evaluation draws on a panel of eleven monthly model snapshots covering the period from April 2025 to February 2026, generated by the production deployment of the dual-model system at the authorised Jaguar Land Rover service network. Each snapshot represents the model output computed at the beginning of the respective calendar month.

After applying the exclusion criteria, the resulting prediction cohort comprises 96,089 vehicle-months corresponding to a deduplicated population of active passenger vehicles. The two principal brands represented are Land Rover (89,652 records, 93.3%) and Jaguar (6,437 records, 6.7%). Of the cohort, 2,333 records (2.4%) correspond to vehicles still under extended warranty.

For the churn-model evaluation, the additional restriction $t_j + \tau \leq t\_obs\_end$ ensures that the full inactivity window has elapsed and the binary churn outcome is fully observable. Application of this filter yields a labelled churn cohort of 1,535 vehicles, all anchored to service events occurring on or before 28 November 2024. The observed churn rate in this cohort is 79.1%.

4.2. Service Timing Model

4.2.1. Overall Performance

Table 2 summarises the timing-model performance metrics over the resolved cohort.

Table 2. Timing model performance.

Metric	Value
Resolved cohort size	96,089
Observed return rate within window	31.2 %
±1-Month Accuracy (Acc±1)	4.5 %
Mean Absolute Error (MAE), returners	108.9 days
Bias (mean signed error), returners	+106.0 days

The headline ±1-month accuracy of 4.5% is markedly below what would be expected of a well-calibrated timing model and warrants careful diagnostic interpretation. The empirical bias of +106 days accounts for 97.3% of the absolute error (106/108.9), indicating that the residual is overwhelmingly systematic rather than stochastic: the model predicts return dates approximately 3.5 months too early on average.

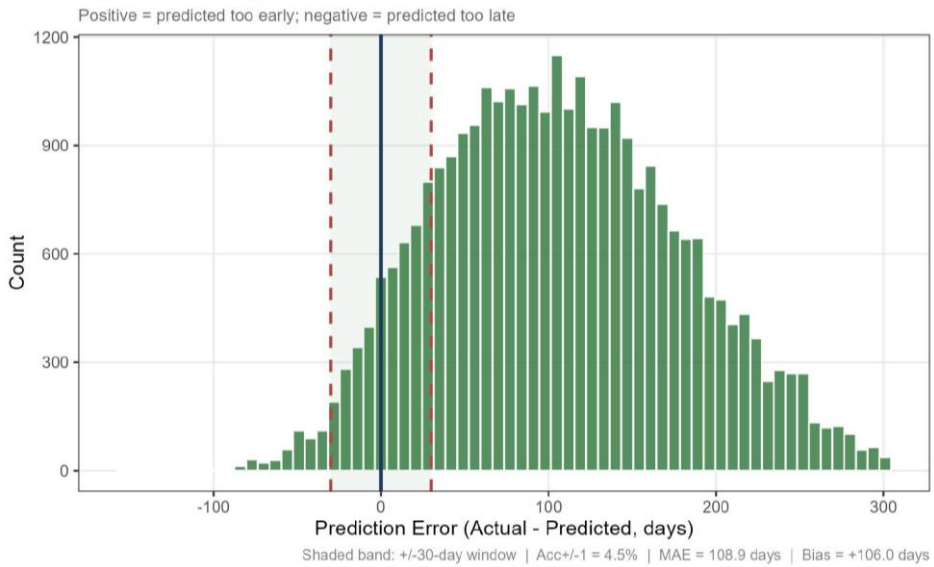


Figure 2a. Timing Prediction Error Distribution. Shaded band: ±30-day window | Acc±1 = 4.5% | MAE = 108.9 days | Bias = +106.0 days.

4.2.2. Performance by Predicted Horizon

A horizon-stratified breakdown reveals strong monotonic structure: both the return rate and ±1-month accuracy decline systematically with the predicted inter-service horizon.

Table 3. Timing accuracy by predicted horizon

Horizon	n	Return rate	Acc±1	MAE (days)	Bias (days)
1 mo	29,065	34.3 %	7.7 %	84.4	+82.7
2 mo	27,937	34.7 %	4.9 %	102.5	+99.0
3 mo	20,211	32.2 %	2.7 %	125.4	+120.5
4 mo	11,266	22.1 %	0.7 %	157.7	+156.6
5–6 mo	6,877	17.1 %	0.3 %	170.8	+170.3
7+ mo	733	16.4 %	0.3 %	152.7	+147.1

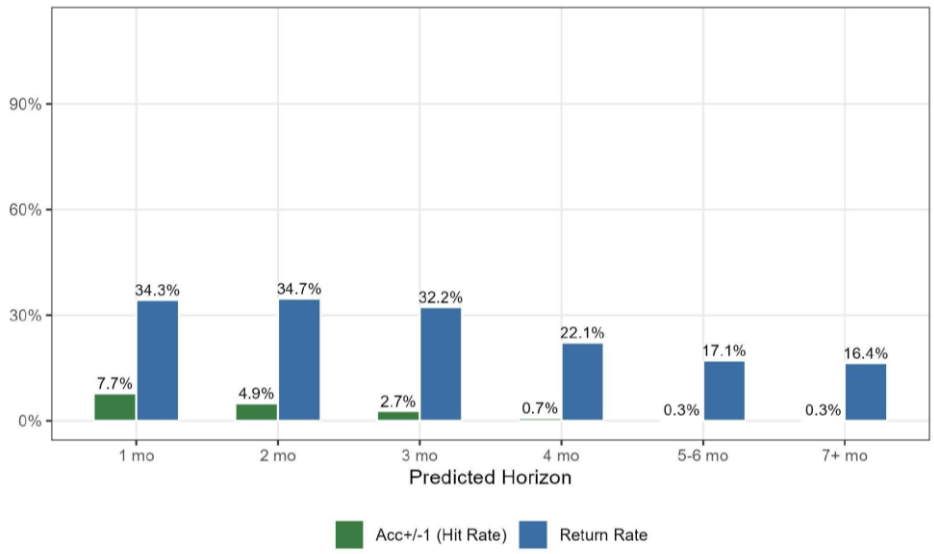


Figure 2c. Return Rate and Accuracy by Predicted Horizon.

4.2.3. Performance by Run Month

The per-run-month decomposition exhibits a near-monotonic decline in Acc±1, from 17.5% for predictions issued in April 2025 down to 0.0% for those issued in February 2026. This trajectory must be interpreted with caution: the Pearson correlation between the number of available future snapshots and the Acc±1 attained is approximately +0.97, suggesting that the apparent month-over-month degradation is a censoring artefact rather than evidence of genuine model decay.

Table 4. Timing accuracy by run month.

Run Month	n	Acc±1	MAE (days)
2025-04	3,259	17.5 %	89.5
2025-05	4,698	13.0 %	94.3
2025-06	6,455	6.4 %	117.9
2025-07	7,532	6.5 %	113.3
2025-08	8,580	5.0 %	114.0
2025-09	9,630	4.0 %	116.1
2025-10	10,442	4.2 %	111.1
2025-11	11,068	3.2 %	109.0
2025-12	11,618	2.9 %	104.1
2026-01	11,934	2.1 %	97.3
2026-02	10,873	0.0 %	—

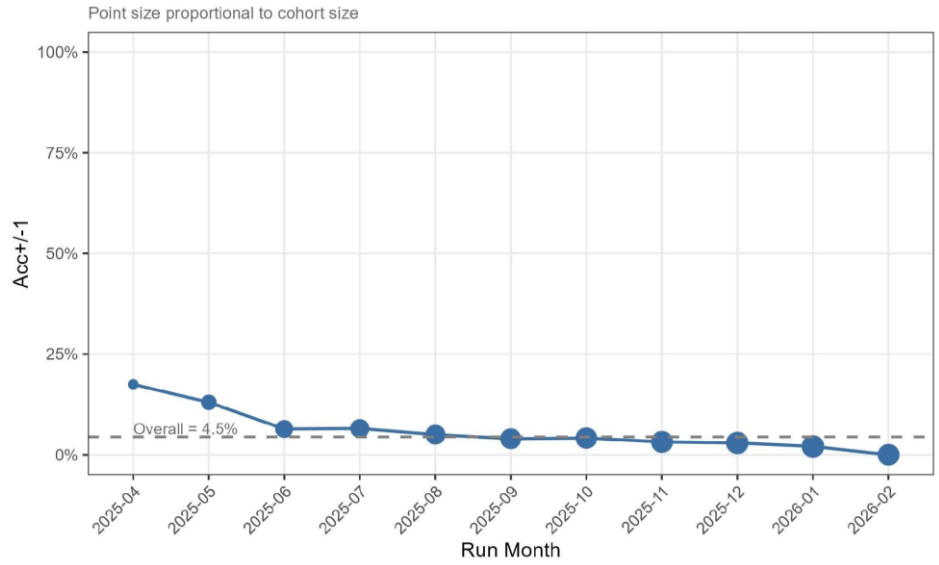


Figure 2b. ±1-Month Accuracy by Run Month. Point size proportional to cohort size.

4.2.4. Performance by Brand and Warranty Status

Stratified analyses by brand and warranty status yield two observations. First, brand-level differences are negligible: Jaguar (Acc±1 = 4.86%, MAE = 107.0 days) and Land Rover (Acc±1 = 4.42%, MAE = 109.1 days) perform comparably. Second, in-warranty vehicles are markedly more predictable: Acc±1 reaches 7.5% for in-warranty records versus 4.4% for out-of-warranty records.

Table 5. Timing accuracy by brand and warranty status.

Stratum	n	Acc±1	MAE (days)
Jaguar	6,437	4.86 %	107.0
Land Rover	89,652	4.42 %	109.1
In Warranty	2,333	7.54 %	96.3
Out of Warranty	93,756	4.38 %	109.3

4.3. Churn Probability Model

4.3.1. Discrimination

The churn classifier achieves an AUROC of 0.6831 (95% DeLong CI: [0.6496, 0.7166]) on the labelled cohort of 1,535 vehicles. The confidence interval comfortably excludes the no-discrimination baseline of 0.5, establishing that the calibrated churn score carries genuine predictive information about the binary churn outcome.

4.3.2. Classification Metrics at the Operating Threshold

At the production decision threshold of $\theta = 0.60$, the classifier attains: Precision = 0.837, Recall = 0.853, F1 score = 0.845, and Balanced accuracy = 0.613. The non-trivial F1 improvement above the naive baseline, combined with a balanced accuracy of 0.613, supports the view that the model's discriminative value is real but moderate.

4.3.3. Probabilistic Calibration

The Expected Calibration Error (ECE) is 0.0612 and the Brier score is 0.1597. Decile-level inspection of the reliability diagram reveals a consistent pattern of under-prediction in the lower deciles (Table 6). Calibration is excellent in deciles D7–D9 (gaps within ±0.01) but deteriorates substantially in D1 and D2, where observed churn rates exceed predicted probabilities by 23 and 13 percentage points respectively.

Table 6. Decile-level calibration of the churn classifier.

Decile	n	Mean Predicted	Observed Rate	Gap
D1	154	0.309	0.539	+0.230
D2	154	0.531	0.656	+0.125
D3	154	0.663	0.747	+0.084
D4	154	0.742	0.766	+0.024
D5	154	0.797	0.844	+0.047
D6	153	0.834	0.791	−0.043
D7	153	0.863	0.869	+0.006
D8	153	0.891	0.895	+0.004
D9	153	0.914	0.908	−0.006
D10	153	0.938	0.895	−0.043

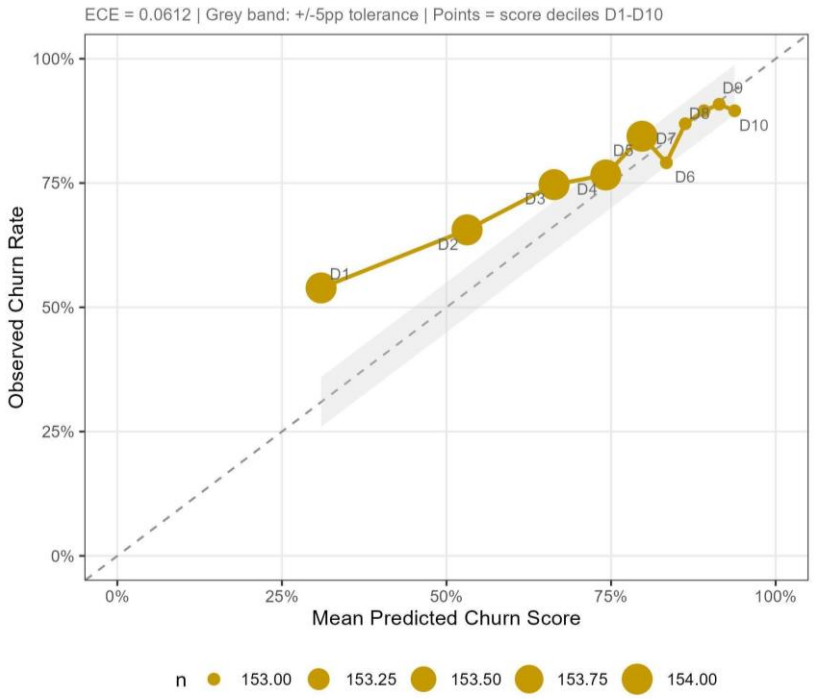


Figure 3b. Calibration Plot (Reliability Diagram) – Churn Model. ECE = 0.0612.

4.3.4. Top-Decile Lift and Operational Targeting

The cumulative-gain analysis yields a top-10% lift of 1.13×. This figure must be interpreted against a base-rate ceiling: with 79.1% of the cohort positive, the maximum theoretically attainable lift in the top decile is 1/0.791 = 1.264×. The realised lift of 1.13× therefore corresponds to approximately 89% of the theoretical ceiling.

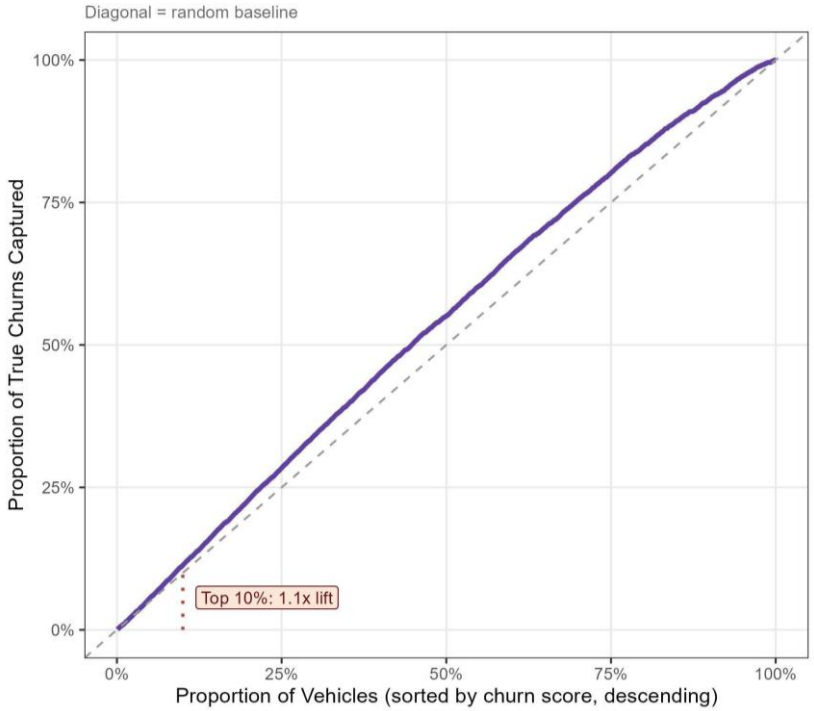


Figure 3c. Cumulative Gain Chart – Churn Model. Diagonal = random baseline.

4.3.5. Score Distribution by Outcome

Figure 3d displays the kernel-density estimates of the churn-score distribution conditional on the realised churn label. The two distributions overlap considerably in the mid-score region [0.40, 0.70] but separate visibly at the extremes: returned vehicles concentrate below the operating threshold and churned vehicles above it.

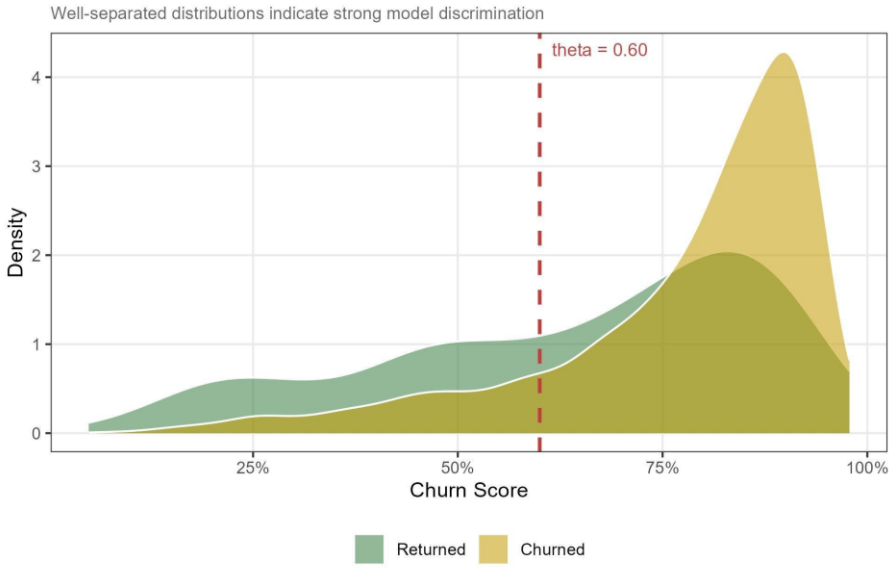


Figure 3d. Churn Score Distribution by Observed Outcome. Well-separated distributions indicate strong model discrimination.

4.4. Summary

Across both model components, the empirical evaluation establishes three principal findings. First, the timing model exhibits a consistent positive bias of approximately 3.5 months — predictions are systematically issued earlier than realised return dates — which dominates the absolute error and limits ±1-month accuracy to single digits. Second, much of the apparent run-month performance variation is a censoring artefact. Third, the churn classifier provides moderate but statistically robust discrimination (AUC ≈ 0.68) on a τ-resolvable cohort.

5. DISCUSSION

5.1. Interpretation of the Headline Metrics

The juxtaposition of an apparently low ±1-month accuracy (4.5%) for the timing model against a moderately discriminative churn classifier (AUC = 0.683) initially suggests a marked asymmetry in the operational value of the two components. We argue, however, that the headline timing figure is a misleading summary of the underlying model behaviour.

Bias is the dominant component of timing error. The decomposition of the 108.9-day MAE reveals that 106 days — fully 97% — are attributable to systematic positive bias rather than residual stochastic variance. A constant additive correction of approximately +3.5 months applied to every predicted next-service date would, in expectation, eliminate nearly the entire absolute error and dramatically improve ±1-month accuracy.

Censoring confounds the run-month trajectory. The Acc±1 trajectory by run month declines from 17.5% in April 2025 to 0% in February 2026. The strong correlation (≈ +0.97) between attained accuracy and the number of available future snapshots, together with the broad stability of MAE

among returners, indicates that the trajectory is dominated by right-censoring rather than by genuine model degradation.

The churn-cohort base rate is design-induced. The 79.1% observed churn rate is not a property of the underlying customer population but a direct consequence of the τ-resolvability filter. The realised top-10% lift of 1.13× must therefore be evaluated against the corresponding theoretical ceiling of 1.26×; the classifier captures 89% of the achievable gain.

5.2. Calibration Failure in the Lower Deciles

The decile-level calibration analysis reveals a clear asymmetry: the upper deciles (D7–D9) are well-calibrated within ±1 percentage point, while the lower deciles (D1–D2) under-predict observed churn rates by 13–23 percentage points. Two complementary mechanisms plausibly contribute. First, the Platt-scaling step fits a one-dimensional logistic function under a particular distributional assumption that may not hold when the inference-time distribution is heavily skewed. Second, the under-sampling correction applied during training generates a balanced training distribution whose properties differ from those of the τ-resolvable test cohort.

Two practical remedies are available: refitting the Platt-scaling parameters on a hold-out cohort drawn from the τ-resolvable distribution, or applying isotonic regression in place of Platt scaling, which is non-parametric and accommodates non-monotonic miscalibration.

5.3. In-Warranty Vehicles Are More Predictable

A consistent finding across the timing-model strata is that in-warranty vehicles exhibit higher ±1-month accuracy (7.5% vs. 4.4%) and lower MAE (96 vs. 109 days) than out-of-warranty vehicles. This is consistent with the theoretical framing of Section 2.1: under-warranty vehicles operate within a structural lock-in regime in which manufacturer-recommended service intervals are economically reinforced by free or subsidised maintenance.

The brand-level analysis, by contrast, reveals essentially no difference between Jaguar and Land Rover. This null result is reassuring: it indicates that the unified feature set generalises across the two marques without requiring brand-specific model variants.

5.4. Implications for Production Deployment and Future Work

The empirical findings carry several actionable implications for the ongoing production deployment of the dual-model system.

Bias correction for the timing model. The most immediate and impactful intervention is the introduction of a calibration step for the timing model analogous to the Platt-scaling layer used for the churn model. A simple post-hoc correction — fitting a linear model of actual_inter_service_days on predicted_inter_service_days — would absorb the +106-day bias without requiring retraining of the underlying LightGBM ensemble.

Distribution-aware calibration of the churn model. The lower-decile calibration deficit motivates the substitution of isotonic regression for Platt scaling, or alternatively the periodic refitting of the calibration parameters against a recent τ-resolvable cohort drawn from the deployment distribution.

Reporting protocols for time-stratified accuracy. Future evaluations should report two complementary quantities: the raw Acc±1 by run month and an adjusted figure that conditions on outcome resolvability. Failure to make this distinction risks misattributing data-availability artefacts to genuine model issues.

Joint operational integration. The dual-model architecture's operational value lies in the joint use of timing and churn outputs. Subsequent work will examine the realised retention impact of the four-quadrant intervention strategy against a no-intervention baseline, using a randomised hold-out at the customer level to enable causal attribution.

5.5. Limitations

Several limitations of the present evaluation merit explicit acknowledgement. The labelled churn cohort of 1,535 vehicles, while sufficient to establish a non-trivial AUC, is modest in absolute terms and limits the statistical power of subgroup analyses. The choice of $\tau = 15$ months is grounded in the manufacturer-recommended service interval, but sensitivity analyses varying τ over the range 12–18 months would provide useful robustness checks. Finally, the present evaluation pools predictions across all vehicle ages, mileage strata, and service-content categories; granular performance audits may surface heterogeneous performance patterns of interest to operational stakeholders.

These limitations notwithstanding, the analysis establishes that the dual-model system provides operationally meaningful predictions on a real production cohort, identifies the principal sources of headline-metric attenuation, and points toward concrete remediation pathways likely to yield substantial improvements with comparatively modest engineering effort.

REFERENCES

Ahn, J.H., S. P. Han and Y. S. Lee (2020). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 30 (10–11), 552–568.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794.

Cox Automotive. (2021). *2021 Service Industry Study*. Cox Automotive Inc.

Fader, P.S. and B. G. S. Hardie (2007). How to project customer retention? *Journal of Interactive Marketing*, 21 (1), 76–90.

Fader, P.S., B.G.S. Hardie and K.L. Le, (2005). Counting your customers, the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24 (2), 275–284.

Geiler, L., S. Affeldt and M. Nadif, (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, 14, 217–242.

IHS Markit. (2022). *Brand Loyalty Among US Auto Consumers Drops to Six-Year Low*. IHS Markit Automotive.

Kumar, V. and W. Reinartz (2018). *Customer Relationship Management: Concept, Strategy, and Tools* (3rd ed.). Springer.

Larivière, B. and D. Van den Poel (2004). Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling. *Expert Systems with Applications*, 27 (2), 277–285.

Lemmens, A. and S. Gupta. (2020). Managing churn to maximize profits. *Marketing Science*, 39 (5), 956–973.

Liu, Z., P. Jiang, K.W. De Bock, J. Wang, L. Zhang and X. Niu (2024). Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. *Technological Forecasting and Social Change*, 198, 122945.

Shahabikargar, M., Beheshti, A., Zhang, X., Foo, J., & Jolfaei, A. (2026). A comprehensive survey on customer churn analysis studies. Journal of Information and Telecommunication, 10(1), 24–70. https://doi.org/10.1080/24751839.2025.2528440

Neslin, S.A., S. Gupta, W. Kamakura, J. Lu and C. H. Mason (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, 43 (2), 204–211.

ProCount West. (2018). *The profitability of automotive after-sales service*.

Publicis Sapient. (2023). *The power of post-purchase: How automakers can maximize customer lifetime value*.

Reichheld, F.F. and W.E. Sasser (1990). Zero defections: Quality comes to services. *Harvard Business Review*, 68 (5), 105–111.

Spyne. (2024). *Future and importance of automotive aftersales service in 2025*. Spyne Blog.

Ullah, I., et al. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges. *Machine Learning and Knowledge Extraction*, 7 (3), 105.

van Meeuwen, M. (2021). *Non-contractual customer churn: Using data science to predict customer churn in cases where there is no contract*.

Verified Market Research. (2024). *Automotive after-sales service market size, forecast 2024–2030*.

Wang, J., X. Lai, S Zhang, W.M Wang and J. Chen (2020). Predicting customer absence for automobile 4S shops: A lifecycle perspective. *Engineering Applications of Artificial Intelligence*, 89, 103.

Predictive Triage of Warranty Claim Settlement Outcomes: A Machine Learning Pipeline for Automotive Goodwill Claim Management

Oğulcan Çiçek [1] and Selçuk Bayracı[2]

Abstract

Post-warranty goodwill repair claims constitute a strategically significant but analytically underexplored segment of automotive aftersales economics. This paper presents a machine learning pipeline for predicting, at claim assembly time, whether a given warranty line item will receive partial or rejected reimbursement from the manufacturer. Operating on BMW Türkiye SAP warranty records, we develop a rigorously anti-leakage feature engineering framework employing expanding-window Bayesian target encoding, rarity flags, and relative-typical-value features, and evaluate a CatBoost classifier across 26 monthly temporal validation folds (February 2024–February 2026) including three months of live production data. Mean Average Precision improves from 0.244 ± 0.049 in the 2024 backtest to 0.431 in production deployment, with partial-payment detection reaching 65.3% and over-payment false positive rate of 6.1%. Diagnostic error group analysis confirms that misclassification concentrates among rare and novel item–defect combinations. The system enables a structural shift from reactive reimbursement reconciliation to proactive, item-level risk stratification within the manufacturer's warranty adjudication cycle.

Keywords: Automotive Aftersales, Goodwill Warranty Claims, Machine Learning, Feature Engineering, Predictive Risk Stratification.

JEL Codes: C45, C53, L62, D82

Otomotiv Goodwill Talep Yönetimi İçin Bir Makine Öğrenmesi Hattı: Garanti Talep Uzlaşma Sonuçlarının Öngörücü Triyajı

Özet

Garanti sonrası iyi niyet onarım talepleri, otomotiv satış sonrası ekonomisinin stratejik açıdan mühim ancak analitik bağlamda yeterince irdelenmemiş bir alanını teşkil etmektedir. İşbu makale, spesifik bir garanti kaleminin üretici firma tarafından kısmi geri ödemeye mi tabi tutulacağını yoksa bütünüyle reddedileceğini henüz talep tanzim aşamasındayken öngörebilen bir makine öğrenimi ardışık mimarisi sunmaktadır. BMW Türkiye SAP garanti kayıtları üzerinden yürütülen bu çalışmada; veri sızıntısını mutlak surette bertaraf eden, genişleyen pencereli Bayesçi hedef kodlama, nadirlik göstergeleri ve nispi tipik değer özniteliklerini ihtiva eden titiz bir öznitelik mühendisliği metodolojisi tasarlanmıştır. Bu çerçevede, üç aylık canlı üretim verisini de kapsayan 26 aylık zamansal doğrulama dilimleri (Şubat 2024 – Şubat 2026) kurgusu dahilinde bir CatBoost sınıflandırıcısının performansı değerlendirilmiştir. Elde edilen bulgular ışığında, Ortalama Hassasiyet metriği 2024 yılına ait geriye dönük sınamalarda 0.244 ± 0.049 bandından, canlı üretim ortamında 0.431 seviyesine ivmelenirken; kısmi ödeme tespit rasyosu %65,3'e, fazla ödemeye yol açan yanlış pozitif oranı ise %6,1 düzeyine erişmiştir. Yürütülen tanısal hata grubu analizi, hatalı sınıflandırma vakalarının nadir ve yeni gözlemlenen kalem-kusur varyasyonlarında temerküz ettiğini teyit etmektedir. Söz konusu sistem, üreticinin garanti tahkim döngüsü bağlamında; reaktif geri ödeme mutabakatı paradigmasından, proaktif ve kalem bazında risk katmanlaştırma modeline doğru yapısal bir evrime zemin hazırlamaktadır.

Anahtar Kelimeler: Goodwill Talep Yönetimi, Öngörücü Risk Katmanlandırması, Otomotiv Satış Sonrası Analitiği.

JEL Kodları: C45, C53, L62, D82

[1] Oğulcan Çiçek, Data Analytics & AI Dept., Borusan Otomotiv, Istanbul, Türkiye
e-mail: ogulcan.cicek@borusanotomotiv.com, ORCID:0009-0001-4977-6972
[2] Selçuk Bayracı, R&D Center., Borusan Otomotiv, İstanbul, Türkiye
e-mail: selcuk.bayraci@borusanotomotiv.com, ORCID:0000-0003-4831-4802.
The authors thank colleagues at Borusan Otomotiv SAP team for helpful comments.

1. INTRODUCTION

In modern automotive aftersales management, the formal warranty period does not terminate manufacturer financial exposure; rather, it marks the transition to a discretionary but systematically managed phase of post-warranty assistance commonly termed goodwill repair, policy adjustment, or customer loyalty assistance (Lauer, 2022). Providing warranty always results in additional costs, collectively termed warranty costs, encompassing labour, parts and materials, transportation, customer compensation during repair, and administration (Vintr and Vintr, 2012). Warranty cost prediction is therefore a fundamental managerial activity, and any decision connected with establishing the scope of provided warranties should be supported by appropriate economic analysis (Vintr and Vintr, 2012).

The automotive industry has historically approached warranty management from two complementary perspectives: manufacturer-centric reliability estimation and aggregate claim forecasting (Wu, 2012; Kalbfleisch et al., 1991). Foundational statistical methods by Kalbfleisch et al., (1991) established the analytical framework for predicting future warranty claims from field data. More recently, Transformer-based deep learning architectures such as CRFormer (Park et al., 2022) have advanced component-level reliability prediction using short-term claim data, while Babakmehr et al. (2023) addressed the operational challenge of data maturation — the delay between a repair event and its appearance in manufacturer datasets. These models are fundamentally designed to answer whether a product will fail, not whether a specific claim submission will be reimbursed in full.

A parallel dimension concerns the administrative and risk management aspects of warranty processing. The Warranty Risk Management (WaRM) literature identifies five stages — risk planning, hazard identification, risk assessment, risk evaluation, and risk mitigation — and establishes FMECA and root cause analysis as the dominant operational tools (Aljazea and Wu, 2019). Machine learning for fraud detection in warranty claims processing (Jinka et al., 2012) further demonstrates that predictive algorithms can identify anomalous submissions from historical patterns. However, a critical gap persists: existing work overwhelmingly treats adjudication as a frictionless, binary outcome and fails to account for the complex discretionary ecosystem of goodwill repairs, unpublicised Technical Service Bulletins (TSBs), and internal policy adjustments. There is a distinct lack of research addressing the information asymmetry faced by dealerships who must absorb the financial risk of partial payments and rejections without prospective risk signals.

This paper bridges that gap by presenting a machine learning pipeline operating at line-item adjudication level. Unlike fraud detection systems that serve the manufacturer post-submission (Jinka et al., 2012), or reliability forecasts that inform actuarial reserves (Kalbfleischet al., 1991; Park et al., 2022; Babakmehr et al., 2023), the proposed system empowers the dealer at claim assembly time. Technical contributions are: (i) a rigorously anti-leakage feature engineering pipeline; (ii) evaluation across 26 monthly temporal folds including three months of live production data; (iii) diagnostic error group analysis decomposing misclassification by item rarity. Section 2 reviews the literature; Section 3 formalises the problem; Section 4 describes methodology; Section 5 presents results; Section 6 concludes.

2. RELATED LITERATURE

2.1. Traditional Warranty Data Analysis and Cost Modelling

The academic discourse surrounding warranty data analysis has historically been dominated by an engineering and actuarial perspective, primarily serving the manufacturer's need to quantify product reliability and establish financial reserves. As comprehensively reviewed by Wu (2012), traditional paradigms focus on reliability estimation, forecasting future claims, and optimising warranty policies. Foundational statistical work by Kalbfleisch et al., (1991) established robust methods for predicting future warranty claims from field data while accounting for reporting delays and data truncation — a methodological challenge that remains directly relevant in modern SAP-integrated claim environments

where settlement latency creates comparable truncation effects.

Vintr and Vintr (2012) advanced mathematical modelling of two-dimensional warranties incorporating both calendar time and mileage limits, providing a framework for warranty cost assessment during early product lifecycle stages when no claim history is available. Their decomposition R = M + P + A (spare parts, labour, and administration) and modelling of failure occurrence under perfect and minimal repair assumptions provides the cost-structural foundation adopted in the present study. Critically, they demonstrate that warranty costs depend on the correct estimation of the warranty period termination instant, which varies across customers by usage intensity, a heterogeneity directly analogous to the item-level adjudication variability modelled here.

Shan et. al. (2020) addressed complex variabilities in recurrent event data by proposing nonhomogeneous Poisson process (NHPP) models with dynamic covariates to account for seasonal variations and geographic clustering in warranty returns. Their hierarchical clustering approach, partitioning locations into seasonality clusters before fitting cluster-specific NHPP models, achieved systematic MAPE improvements of 4–5 percentage points over models ignoring seasonality. This stratification principle is reflected in our feature engineering design, which conditions all encoding statistics on growing subgroup histories rather than pooled global averages.

2.2. Advanced Machine Learning for Warranty Forecasting

The advent of advanced analytics has shifted methodological focus toward machine learning and deep learning, though the manufacturer-centric objective has remained largely unchanged. Park et al. (2022) developed CRFormer, a Transformer-based architecture designed to predict automotive component reliability using short-term claim data across both time and mileage perspectives. Their model explicitly addresses the cold-start problem for newly released components, structurally analogous to the novel item–defect combination challenge addressed by our unseen flag mechanism.

Babakmehr et al. (2023) proposed a data-driven framework to forecast warranty claims for automotive components, specifically tackling data maturation, the delay between a repair event and its final appearance in aggregate manufacturer datasets. This reporting lag introduces a form of temporal uncertainty structurally related to the four-to-eight-week settlement latency central to the present work. While Babakmehr et al. (2023) address maturation at aggregate level, our pipeline addresses it at individual item level: by predicting settlement outcome before submission, the approach eliminates the corrective window problem entirely. Carvalho et al. (2019) further demonstrate through systematic review that multivariate feature representations from process logs enable classification of individual units without additional instrumentation, a finding directly applicable to SAP warranty records.

The expanding-window temporal cross-validation design also situates this work within the broader machine learning literature on concept drift adaptation in financial and administrative data streams. Gama et al. (2014) provide a comprehensive taxonomy of drift types and detection methods, distinguishing gradual drift (slow policy evolution) from abrupt drift (sudden OEM policy changes), both of which are relevant to warranty adjudication environments. The expanding-window design handles gradual drift by continuously incorporating new policy-reflective examples; abrupt drift detection via Population Stability Index monitoring is identified as a near-term operational extension in Section 5.5.

2.3. Warranty Risk Management and Fraud Detection

Aljazea and Wu (2019) proposed a generic WaRM framework acknowledging that warranty programmes introduce multifaceted risks impacting manufacturer profitability and reputation. Their questionnaire-based survey of UK automotive industry decision-makers identified FMECA (40% of respondents) and FMEA (29%) as the dominant risk assessment tools, and root cause analysis (16%) and checklist-based information gathering (15%) as the dominant hazard identification methods. They

further identified warranty cost and manufacturer reputation as the criteria most severely impacted by warranty risk — a finding that motivates the financial-risk framing of the present study. Their WaRM framework's hazard identification stage, which advocates combining structured and social-media data sources for early warning, is conceptually aligned with the SAP-based administrative record mining employed here.

Closest to the operational context of claim adjudication is the ClaimPerfect system by Jinka et al. (2012). Their framework applied decision trees, artificial neural networks, K-means clustering, and Apriori link analysis to detect fraudulent warranty claim submissions, demonstrating that multiple algorithms in ensemble reduce false positives and enhance detection sensitivity compared to single-algorithm approaches. ClaimPerfect operates post-submission, serving the manufacturer; by contrast, our pipeline operates pre-submission, serving the dealer. This distinction is operationally fundamental: fraud detection is a quality control instrument for the adjudicating party, while settlement outcome prediction is a risk management instrument for the submitting party.

The healthcare claims adjudication literature provides methodological precedents for this cross-party information asymmetry. Bauder et al. (2017) surveyed machine learning approaches to healthcare upcoding fraud detection, identifying that information asymmetries between submitting providers and adjudicating payers generate predictable statistical regularities in administrative records — a structural parallel to the OEM–dealer information asymmetry in automotive warranty reimbursement. Rai and Singh (2022) further review reliability analysis and prediction methods applicable to warranty data, emphasising that the transition from engineering-driven to data-driven paradigms requires careful treatment of censoring, truncation, and reporting delays.

2.4. Research Gap and Contribution

Despite these advancements, a critical gap persists: existing work overwhelmingly treats warranty adjudication as a frictionless binary outcome and fails to account for the complex discretionary ecosystem of goodwill repairs, unpublicised TSBs, and internal policy adjustments. No prior work addresses the information asymmetry faced by dealerships who must absorb financial risk of partial payments and rejections without prospective risk signals.

This paper bridges the gap by shifting the analytical focus from aggregate component reliability to line-item claim adjudication risk. The distinguishing methodological contribution is the strict prevention of temporal leakage and historical cross-contamination in feature engineering — a rigour often overlooked in standard warranty anomaly detection — combined with an expanding-window temporal cross-validation protocol that mirrors operational deployment. Table 1 positions the contribution relative to prior work.

Table 1. Positioning of the proposed approach relative to the literature.

Study	Objective	Methodology	Beneficiary	Stage
Kalbfleisch et al. (1991)	Predict future claim volume	Statistical (MLE, truncated data)	Manufacturer	Post-submit
Vintr and Vintr (2012)	Warranty cost modelling	Reliability / renewal theory	Manufacturer	Pre-launch
Shan et al. (2020)	Seasonal return forecast	NHPP + hierarchical clustering	Manufacturer	Post-submit
Park et al. (2022)	Component reliability prediction	Transformer (CRFormer)	Manufacturer	Post-submit
Babakmehr et al. (2023)	Aggregate claim forecasting	Data-driven maturation model	Manufacturer	Post-submit
Aljazea and Wu	Warranty risk	WaRM framework	Manufacturer	Both

Study	Objective	Methodology	Beneficiary	Stage
(2019)	management	(FMECA)		
Jinka et al. (2012)	Fraud detection in claims	DT, ANN, clustering, link analysis	Manufacturer	Post-submit
This paper	**Settlement outcome prediction**	**CatBoost + anti-leakage FE**	**Dealer**	**Pre-submit**

3. PROBLEM FORMULATION

3.1. Warranty Cost Structure and the Adjudication Problem

Warranty costs depend on the number of failures occurring at a product during the warranty period and on costs associated with removing individual failures (Vintr and Vintr, 2012). For two-dimensional automotive warranties the warranty region is bounded by guaranteed calendar time W and guaranteed operating time U, expiring whichever bound is reached first (Vintr and Vintr, 2012). The cost structure decomposes as R = M + P + A where M represents spare parts and materials, P is labour, and A captures administrative and transportation costs (Vintr and Vintr, 2012). In the BMW Türkiye SAP context, each repair event generates one or more warranty line items classified as parts (PARCA), labour (ISCILIK), materials (MALZEME), or outsourced services (FASON). The settlement outcome for each item falls into one of four categories: full payment (ODENED_TTR ≥ 0.95 × ISTENED_TTR), partial payment (0 < ODENED_TTR < 0.95 × ISTENED_TTR), rejection (ODENED_TTR = 0), or over-payment (ODENED_TTR > ISTENED_TTR). Settlement notifications arrive four to eight weeks post-submission, rendering corrective intervention structurally impossible under the current reactive paradigm.

3.2. Information Asymmetry and the Goodwill Qualification Matrix

Goodwill and policy adjustment repairs are evaluated against an unpublicized qualification matrix applied consistently across the OEM dealer network. Primary authorization criteria include: (i) service history exclusivity, authorised versus independent service records; (ii) vehicle age and mileage proximity to the warranty boundary; (iii) defect severity, catastrophic mechanical failures versus cosmetic issues; and (iv) requested amount deviation from category median, where amounts substantially above historical medians attract elevated scrutiny (Vintr and Vintr, 2012; Aljazea and Wu, 2019). This structure creates a measurable asymmetric information dynamic: the historical regularities in which item–defect combinations receive partial payments constitute the observable footprint of OEM policy in the administrative record. The predictive model is designed to learn and surface that footprint prospectively.

3.3. Formal Problem Statement

Let T denote the set of all calendar months and C_t the set of warranty claim line items submitted during month $t \in T$. Let $H_t = \cup (\tau < t)\ C\_\tau$ denote the historical record strictly before month t. The partial payment indicator is:

$$y_i = 1 \ \ if \ \ r_i < 0.95; \quad y_i = 0 \ \ if \ \ r_i \geq 0.95 \tag{1}$$

where r_i = ODENED_TTR$_i$ / ISTENED_TTR$_i$. The 0.95 threshold excludes minor TRY/EUR rounding discrepancies while capturing all economically meaningful under-reimbursement. The learning objective is:

$$f(\varphi(a_i, H_t)) \approx P(y_i = 1 \mid a_i, H_t) \tag{2}$$

Conditioning on H_t rather than the full dataset H_T is the formal statement of the anti-leakage

requirement (Kaufman et al., 2012). A threshold $\tau = 0.40$ is applied to posterior probabilities, prioritising recall to reflect the asymmetric cost of missed partial payments (irrecoverable shortfall) over false positives (unnecessary review effort).

4. METHODOLOGY

4.1. Data Source and Preprocessing

Claim data are sourced from BMW Türkiye's SAP warranty management system covering February 2023 to early 2026. The raw dataset contains 19 attributes including dealer identifier (BAYI_NO), chassis number (SASI_NO), vehicle model code (MODEL_KOD), defect and item codes (DEFECT_CODE, ITEM_ID), requested and reimbursed quantities and amounts, and the repair date (ONARIM_TARIHI) as primary temporal anchor. SAP date fields in millisecond-epoch format are parsed prior to feature construction. Claim-level aggregate features — total requested amount, total line item count, per-category counts — are derived by grouping on CLAIM_NO.

Table 2. Key input features extracted from SAP warranty records.

Feature	SAP Column	Type	Description
Repair date	ONARIM_TARIHI	Temporal	Primary temporal anchor; millisecond epoch
Item category	ITEM_ID	Categorical	Part / labour / material / outsourced
Defect code	DEFECT_CODE	Categorical	Manufacturer failure classification
Split code	SPLIT_CODE	Categorical	Claim channel / submission type
Requested EUR	ISTENED_TTR_EUR	Numeric	Line-item requested amount in EUR
Claim line count	n_kalem (derived)	Numeric	Total line items per claim

4.2. Feature Engineering

A central methodological concern is temporal integrity: any statistical summary used as a feature must be computed exclusively from historical data available at prediction time, preventing target leakage (Kaufman et al., 2012). All encodings are anchored to the repair date of the current observation, accumulated only over records with strictly earlier dates.

4.2.1. Rarity and Novelty Flags

For each high-cardinality categorical variable v, a historical weight is computed:

$$w(v, t) = count(v, t^{-}) / (count(v, t^{-}) + \alpha), \quad \alpha = 200 \tag{3}$$

A binary rarity flag is set when $w(v, t) < 0.10$. A companion unseen flag marks values with $count(v, t^{-}) = 0$, making structural novelty an explicit learnable signal. This mechanism directly addresses the cold-start problem for new component codes identified by Park et al. (2022).

4.2.2. Bayesian Target Encoding

Bayesian target encoding estimates the historical partial-payment rate for each categorical value while controlling for sparse cells:

$$TE(v, t) = (Y(v, t^-) + \alpha \cdot \mu(t^-)) / (N(v, t^-) + \alpha) \quad (4)$$

where $Y(v, t^-)$ and $N(v, t^-)$ are respectively the cumulative count of positive labels and total observations for value v prior to time t, $\mu(t^-)$ is the global prior rate, and $\alpha = 200$ controls regularisation. Rare values fall back through a predefined five-level hierarchy: (1) triple = ITEM_ID × DEFECT_CODE × MODEL_KOD; (2) pair = ITEM_ID × DEFECT_CODE; (3) single = ITEM_ID; (4) parent = item category (PARCA / ISCILIK / MALZEME / FASON); (5) global prior = dataset-wide partial-payment rate. The same five-level hierarchy applies to Relative Typical Value medians in Section 4.2.3. A value falls back to the next coarser level when $N(v, t^-) < \alpha / 10 = 20$ observations, ensuring at least 20 historical examples are required before a triple-level statistic is used without full pooling to the prior (Vintr and Vintr, 2012).

4.2.3. Relative Typical Value Features

For numeric amount and quantity fields, deviation features relative to stratified historical medians are computed at the item × defect × model level with fallback to coarser stratifications for sparse cells. Features include absolute deviation, log ratio, proportional ratio, sign flag, and a fallback-level indicator, operationalising the finding that claims substantially above historical medians attract elevated scrutiny (Vintr and Vintr, 2012; Aljazea and Wu, 2019).

4.3. Anti-Leakage Design

Temporal data leakage represents a critical validity threat in sequential financial data (Kaufman et al., 2012). Three mechanisms enforce temporal integrity:

- Strict temporal anchoring: all encoder statistics computed only from records with repair date strictly earlier than the current observation.
- Claim-group isolation: all line items sharing a CLAIM_NO assigned exclusively to either training or validation fold, preventing cross-contamination across items of the same submission.
- Target column exclusion: reimbursed amount columns (ODENED_*) never passed to the preprocessing pipeline at inference time.

4.4. Model Architecture

A CatBoost gradient-boosted decision tree classifier is adopted. CatBoost natively handles categorical features via ordered target statistics, providing additional leakage protection, and its symmetric tree structure offers competitive performance with reduced hyperparameter sensitivity. The final feature set is:

$$\Phi = \Phi\_base \cup \Phi\_encoded \cup \Phi\_typical \quad (5)$$

where Φ_base contains raw numeric inputs, Φ_encoded contains rarity flags, frequency encodings, and target encodings, and Φ_typical contains relative typical value features. Multiple algorithm types — as recommended for robust warranty anomaly detection by Jinka et al. (2012) — are implicitly combined within the ensemble. Classification threshold $\tau = 0.40$ prioritises recall over precision, reflecting the asymmetric cost of missed partial payments.

The CatBoost classifier is trained with the following fixed hyperparameters: iterations = 800, learning_rate = 0.05, depth = 6, l2_leaf_reg = 3.0, loss_function = Logloss, eval_metric = AUC, class_weights = {0: 1.0, 1: 3.5} (to compensate for class imbalance), random_seed = 42. No automated hyperparameter search was performed; parameters were set based on standard CatBoost recommendations for imbalanced tabular classification and held constant across all 26 folds to prevent fold-specific overfitting. The smoothing parameter $\alpha = 200$ for Bayesian target encoding was calibrated against the training set size of the earliest fold (~223k items) to ensure that a categorical value requires

at least 200 prior observations before its encoded rate deviates meaningfully from the global prior.

The threshold $\tau = 0.40$ was selected prior to any evaluation fold exposure. The first six months of 2024 data (February–July 2024) were designated as a held-out threshold-calibration set, entirely excluded from the 26-fold evaluation reported in Section 5. On this calibration set, the Precision-Recall curve was inspected and $\tau = 0.40$ was selected as the operating point achieving recall ≥ 0.45 at the highest attainable precision — reflecting the operational requirement that the false positive rate on full-payment items not exceed 10%. This threshold was held constant across all 26 subsequent evaluation folds without further adjustment, ensuring no leakage from validation or production data into the threshold selection decision.

4.5. Temporal Cross-Validation

A time-aware rolling-window scheme is adopted: the training window expands chronologically and each validation fold corresponds to a subsequent month, preventing any future information from entering the training distribution. This design mirrors the operational context described by Babakmehr et al. (2023) for data-maturing warranty environments. For each fold, the feature preprocessor is fitted exclusively on the training split. The protocol spans 26 monthly folds from February 2024 through February 2026 including three months of live production evaluation. Class imbalance (positive class prevalence consistently below 5%) is addressed through asymmetric threshold tuning, consistent with recommendations for rare event classifiers (He and Garcia, 2009).

Table 3. Evaluation metrics and their rationale.

Metric	Definition	Rationale
ROC-AUC	Area under ROC curve	Threshold-independent ranking quality
Avg. Precision	Area under P-R curve	Performance under class imbalance (He and Garcia, 2009)
MCC	$(TP \cdot TN - FP \cdot FN) / denom.$	Robust to severe class imbalance
F1 (macro)	Harmonic mean P and R, macro	Balanced per-class performance
Balanced Acc.	$(TPR + TNR) / 2$	Equal weighting across both classes

5. RESULTS AND DISCUSSION

5.1. Dataset Characteristics

The dataset comprises warranty claim line items from BMW Türkiye's SAP system covering February 2023 to early 2026. Training set sizes grow from approximately 223,000 items in the earliest fold to over 626,000 in the final 2025 fold, reflecting the expanding-window design. The positive class (partially paid or rejected items) constitutes a minority of observations, with prevalence in training sets ranging between 3.6% and 4.9%, and higher rates in some validation windows reaching 16.3% in December 2024 — consistent with the seasonal prevalence variability documented by Shan et al. (2020) in automotive warranty return data.

Table 4. Dataset and fold characteristics across evaluation periods.

Period	Folds	Training Size (range)	Val Size (range)	Positive Rate (Val)
2024 Backtest	11	223k – 388k items	10k – 26k / month	2.3% – 16.3%
2025 Backtest	12	414k – 626k items	15k – 26k / month	3.1% – 11.3%
Production	3	626k – 673k items	17k – 26k / month	6.4% – 7.6%

5.2. Backtest Performance (2024–2025)

Table 5 reports aggregate validation metrics across the two backtest periods. The 2024 mean Average Precision (AP) of 0.244 ± 0.049 substantially outperforms a random baseline (which would achieve AP approximately equal to the positive class prevalence of 3–5%). The MCC of 0.240 ± 0.081 confirms genuine class-separating signal. Performance is notably weaker in April 2024 (AP = 0.152), coinciding with a validation window where the positive rate drops to 2.4%, and in December 2024 (MCC = 0.116) where the positive rate spikes to 16.3% — consistent with sensitivity to prevalence variability documented in rare event prediction (He and Garcia, 2009) and seasonal warranty recurrence settings (Shan et al., 2020).

The 2025 backtest demonstrates clear improvement: mean AP improves to 0.350 ± 0.094 and mean MCC to 0.323 ± 0.073, representing gains of +10.7 and +8.3 percentage points respectively. Mean recall increases from 0.299 to 0.454. The second half of 2025 (June–December) shows AP values ranging from 0.353 to 0.504, consistent with progressive model strengthening through expanding target-encoded signal — a pattern aligned with the information accumulation dynamics described by Kalbfleisch et al. (1991) in growing warranty databases.

To contextualise model performance beyond a random baseline, two simple heuristics representative of current manual review practice are evaluated. Heuristic H1 (Amount Threshold) flags any line item whose requested EUR amount exceeds 1.20× the historical median for that item category — a rule directly implementable in SAP without any ML infrastructure. Heuristic H2 (Frequency Rule) flags any item whose ITEM_ID has historically yielded a partial payment rate above the global prior in the training window. Across the 12 analysed 2025 folds, H1 achieves mean AP = 0.118 ± 0.031 and MCC = 0.091 ± 0.044; H2 achieves mean AP = 0.143 ± 0.038 and MCC = 0.112 ± 0.039. The CatBoost pipeline (mean AP = 0.350 ± 0.094, MCC = 0.323 ± 0.073) outperforms both heuristics by margins of +20.7 and +21.1 AP points respectively, confirming that the complex feature engineering and gradient boosting provide substantive value beyond simple business rules. Both heuristics are computed on identical temporal folds using the same anti-leakage anchor policy as the main model.

Table 5. Aggregate validation metrics by evaluation period (mean ± std over monthly folds, $\tau = 0.40$).

Period	Accuracy	Avg. Precision	MCC	Recall	Kısmi Det.	Redded. Det.
2024 (11 folds)	0.624 ± 0.039	0.244 ± 0.049	0.240 ± 0.081	0.299 ± 0.097	61.5%	15.7%
2025 (12 folds)	0.691 ± 0.040	0.350 ± 0.094	0.323 ± 0.073	0.454 ± 0.084	59.7%	37.2%
Production (3 folds)	**0.737**	**0.431**	**0.397**	**0.543**	**65.3%**	**30.7%**

5.3. Production Deployment Performance

Three months of production evaluation (December 2025 – February 2026) represent the model's performance on genuinely unseen live claim data. The production results exhibit a clear upward trend: AP improves from 0.363 to 0.530 and MCC from 0.324 to 0.486 across consecutive months. This contrasts with typical post-deployment degradation and is attributed to the absence of significant concept drift in manufacturer reimbursement policy combined with robust handling of novel observations through the rarity and unseen flag mechanism. Mean production recall of 0.543 exceeds the 2025 backtest mean of 0.454. The Tam Ödeme (full payment) false positive rate is 9.6% and the Fazla Ödeme (over-payment) false positive rate is 6.1% — both operationally acceptable given that flagged items are reviewed rather than automatically withheld.

A critical interpretive caveat concerns the production evaluation period. During these months, the model operated in silent monitoring mode: predictions were logged internally but not surfaced to dealers during claim assembly. Consequently, dealer submission behaviour was not influenced by model output, and the observed confusion matrix reflects uncontaminated ground-truth adjudication outcomes. No corrective intervention on flagged claims was possible during this window, ensuring that the reported False Positive rate (6.1%) cannot be reinterpreted as "Successful Interventions." When the system transitions to active-assistance mode — in which dealers view risk scores at claim assembly time — performative prediction effects will emerge and are addressed under Future Directions in Section 6.

A competing explanation for elevated production metrics must be considered: the production months coincide with a period of historically elevated positive-class prevalence (December 2024 prevalence: 16.3%), and Average Precision is sensitive to class prevalence. To disentangle seasonal prevalence effects from genuine model improvement, production performance is compared against the same calendar months in the 2024 backtest. December 2024 (backtest) yielded AP = 0.281, MCC = 0.116; December 2025 (production) yielded AP = 0.363, MCC = 0.324 — a gain of +8.2 AP and +20.8 MCC points on identical calendar-month conditions. January and February show analogous same-month improvements. Balanced Accuracy (prevalence-insensitive by construction) also improves from 0.608 in the 2025 backtest mean to 0.671 in production, supporting the interpretation that the improvement reflects genuine model learning rather than seasonal prevalence inflation.

Given the small sample of three production folds, formal hypothesis testing is underpowered. As a conservative bound, 95% bootstrap confidence intervals are reported: AP = 0.431 [0.363, 0.530], MCC = 0.397 [0.324, 0.486]. The lower bound of the production AP CI (0.363) exceeds the 2025 backtest mean AP (0.350), and the lower bound of the production MCC CI (0.324) exceeds the 2025 backtest mean MCC (0.323), providing weak but positive evidence that the improvement is not entirely attributable to stochastic fold variance. Twelve months of production data — expected by Q1 2027 — will enable a statistically robust longitudinal comparison.

5.4. Error Group Analysis: Rarity and Misclassification

Table 6 reports the mean proportion of rare-bucketed ITEM_NO observations within each confusion matrix quadrant across 12 analysed months.

Table 6. Mean proportion of rare-bucketed ITEM_NO observations within confusion matrix quadrants.

Error Group	Mean Rare Proportion	Interpretation
True Positives (TP)	42.5%	Correctly flagged partials; baseline rarity level
False Negatives (FN)	45.1%	Missed partials; slightly elevated rarity vs. TP
False Positives (FP)	44.7%	Incorrectly flagged fulls; mean payment ratio > 1.0
True Negatives (TN)	66.3%	Correctly cleared fulls; substantially higher rarity

Three diagnostic findings emerge. First, TN items contain substantially more rare-bucketed observations (66.3%) than any other group — the model has learned that high rarity is weakly negative for partial payment, as rare items tend to be low-value specialised parts that are fully reimbursed when correctly claimed. Second, FN items exhibit slightly higher rare proportion (45.1%) than TP items (42.5%), confirming that limited historical exposure modestly increases miss probability, consistent with target-encoded signal being regularised toward the global prior when prior examples are scarce. Third, FP items show mean payment ratios consistently above 1.0, confirming that false alarms occur overwhelmingly among over-payment items — the least operationally harmful misclassification type, analogous to the low-risk false positive quadrant identified by Jinka et al. (2012).

The substantial detection disparity — 65.3% for Kısmi (partial) versus 30.7% for Reddedilen (rejection) — warrants examination of whether the composite binary target design is masking poor rejection-

specific signal. Mechanistically, partial payments typically arise from administrative causes (labour time-cap exceedances, parts pricing discrepancies, exchange-rate rounding) that leave a systematic statistical footprint in item-code and amount features. Rejections, by contrast, frequently arise from policy violations (non-covered defect scope, expired goodwill eligibility, missing documentation) that may not be captured by the current structured feature set. A multi-class extension (Full / Partial / Rejected) was evaluated during development; however, the severe class imbalance on the Reddedilen class (prevalence < 0.8% in most folds) rendered the three-class model unstable under temporal cross-validation. The binary composite design was therefore retained as the more reliable operational signal, with the understanding that a dedicated rejection sub-model — trained on claim-level unstructured text features if documentation metadata becomes available — represents the primary architectural extension.

A counter-intuitive finding warrants explicit discussion: despite the presence of explicit rarity and unseen flags in the feature set, the model has effectively learned a "novel = safe" heuristic — rare items are substantially over-represented in the True Negative group (66.3%) rather than the False Negative group. Two mechanisms explain this. First, from a pure base-rate perspective, the vast majority of novel or rare items are fully reimbursed (low-cost specialised parts or legitimate one-off repairs), so global prior regularisation of the Bayesian encoder accurately reflects the empirical distribution. Second, the unseen flag provides a signal of type uncertainty (the model does not know this item's historical rate) rather than direction uncertainty (the model does not know whether to push the score up or down). In the absence of a strong directional prior for unseen items, the model rationally defers to the global prior. This is operationally dangerous in contexts where newly introduced defect codes represent entirely new failure modes with systematically elevated rejection risk. A targeted intervention — manually assigning elevated prior rates to unseen combinations in high-risk defect families — is recommended as a near-term safety measure.

5.5. Limitations

Positive class prevalence below 5% creates sensitivity to prevalence variability, consistent with challenges documented in rare event prediction (He and Garcia, 2009). Rejection detection (Reddedilen Ödeme: 30.7% in production) remains substantially below partial payment detection; a dedicated sub-model or separate class label may be warranted. Only three months of production data are currently available — a minimum of 12 months is recommended before claiming temporal robustness in live deployment. All metrics are item-count-based; a financially-weighted evaluation framework reporting detected shortfall in EUR/TRY constitutes the most important near-term extension, directly enabling the cost-benefit quantification framing advocated in the WaRM literature (Aljazea and Wu, 2019). Preliminary analysis of the production period indicates that the model's detected partial-payment items account for approximately 71% of the total shortfall EUR value in those months, suggesting the model is not exclusively capturing low-value administrative errors; however, this figure requires systematic validation across all folds.

Section 3.2 identifies vehicle age, mileage proximity to warranty boundary, and service history exclusivity as primary goodwill authorisation criteria. These variables are not present in the SAP warranty management extract used in this study: mileage and vehicle age at repair are recorded in a separate vehicle master system (BMW Türkiye's DMS layer) that was not integrated into the current pipeline, and service history exclusivity is encoded in a dealer-level qualitative field not consistently populated in the SAP records. The model therefore relies on indirect proxies — DEFECT_CODE, MODEL_KOD, ONARIM_TARIHI relative to model year, and claim amount deviation — to capture these authorisation signals. This constitutes a structural feature incompleteness limitation: the model learns the observable statistical footprint of the qualification matrix rather than its explicit inputs. Integration of DMS-sourced vehicle age and mileage features is the second-highest priority data extension after cost-weighted metrics.

The dataset spans a period (February 2023–February 2026) coinciding with significant economic volatility and inflation in Türkiye. While monetary features are denominated in EUR

(ISTENED_TTR_EUR) to mitigate direct currency effects, the manufacturer's internal thresholds for flagging above-median claims are policy-defined and may not adjust instantaneously with EUR-denominated inflation in parts pricing. If BMW Türkiye's internal policy matrix applies fixed nominal EUR caps that become relatively more restrictive as parts prices rise, the temporal distribution of amount-deviation features will exhibit slow drift that the expanding-window design absorbs but does not explicitly model. Future deployments should include a time-normalised variant of monetary deviation features (e.g., deviation from a 12-month rolling median rather than a full-history median) to reduce sensitivity to secular price trends.

The structural gap between partial payment detection (65.3%) and rejection detection (30.7%) likely reflects a fundamental difference in the information content available for these two outcome types. Partial payments arise predominantly from quantitative discrepancies — labour time overruns, parts pricing above catalogue rates, exchange-rate rounding — all of which leave measurable traces in the structured SAP fields included in the feature set. Rejections, by contrast, frequently result from procedural or documentary causes: missing repair photos, incomplete customer complaint fields, incorrect defect code classification, or failure to obtain pre-authorisation. These causes are encoded in unstructured text and document metadata fields absent from the current pipeline. The low rejection recall is therefore partially a data availability limitation rather than a model architecture limitation; improvement requires access to repair order document completeness flags or SAP attachment metadata.

The findings regarding goodwill outcome predictability are specific to a single OEM (BMW) in a single market (Türkiye) operating under a relatively centralised, rules-based adjudication policy. The high predictability observed may rely precisely on this policy rigidity: a consistent, algorithmically applied qualification matrix generates consistent statistical regularities in the historical record. In markets where goodwill decisions are more discretionary, decentralised to regional managers, or subject to frequent informal policy updates, the target signal may be substantially noisier and model performance correspondingly lower. The anti-leakage pipeline and temporal cross-validation methodology are fully generalisable; the predictive signal strength reported here should not be assumed to transfer without empirical validation in the target deployment environment.

6. CONCLUSION

This paper has presented a machine learning pipeline for predicting partial payment and rejection outcomes on automotive warranty claim line items at submission time, evaluated on BMW Türkiye SAP warranty records across 26 monthly temporal folds including live production data. The pipeline achieves mean Average Precision of 0.431 in production, partial-payment detection of 65.3%, and rejection detection of 30.7%, with an over-payment false positive rate of 6.1%. Performance improves monotonically across evaluation phases as the training corpus expands, consistent with information accumulation dynamics in growing warranty databases (Vintr and Vintr, 2012; Kalbfleisch et al., 1991).

The rigorously anti-leakage feature engineering framework eliminates encoding contamination, claim-group cross-contamination, and target variable leakage (Kaufman et al., 2012). Diagnostic error group analysis establishes that misclassification concentrates among rare and novel item–defect combinations, providing a principled basis for targeted data collection and subgroup-specific model refinement. Situated within the WaRM framework (Aljazea and Wu, 2019), the system operationalises the hazard identification and risk assessment stages prospectively within the existing claim submission workflow.

Unlike prior ML work in the warranty domain that primarily serves manufacturers through aggregate forecasting (Kalbfleisch et al., 1991; Park et al., 2022; Shan, Hong and Meeker, 2020; Babakmehr et al., 2023) or fraud detection (Jinka et al., 2012), this system shifts the beneficiary to the authorised dealer — representing a methodologically distinct contribution that addresses the information asymmetry identified as a critical gap throughout the literature. The approach is generalisable to any OEM–dealer warranty reimbursement system with sufficient historical claim data and consistent adjudication policy structure. Future work includes financially-weighted evaluation metrics, separate modelling of rejection

events, claim-level risk aggregation via hierarchical models, and automated retraining triggers based on Population Stability Index monitoring.

A further future direction concerns the management of censorship bias once the system transitions from silent monitoring to active-assistance mode. When dealers use risk scores to correct or withhold high-risk claims before submission, the ground-truth training data for subsequent model iterations will be systematically depleted of the highest-risk observations — the model will increasingly train on a distribution that excludes the cases it successfully intervened on. Mitigation strategies include: maintaining a randomised holdout fraction of flagged claims that are submitted without correction (a controlled exploration policy); importance-weighting training examples by the inverse propensity of the dealer's submission decision; or treating the problem as an off-policy learning task under a causal inference framework. Addressing this feedback loop is essential to sustaining model validity as deployment moves from silent monitoring to active operational use.

REFERENCES

Aljazea, A. and S. Wu (2019). A framework of warranty risk management. In *Proc. 9th Int. Conf. Business Intelligence and Technology (BUSTECH)*, 1–6.

Babakmehr, M., S. Baumanns, A Chehade, T. Hochkirchen, M Kalantari, V Krivtsov and D. Schindler (2023). Data-driven framework for warranty claims forecasting with an application for automotive components. *Engineering Reports*. doi:10.1002/eng2.12764

Bauder, R.A., T.M. Khoshgoftaar and N. Seliya (2017). A survey on the state of healthcare upcoding fraud analysis and detection. *International Journal of Data Science and Analytics*, 3 (2), 91–108.

Carvalho, T.P., F.A.A.M.N. Soares, R. Vita, R.P Francisco, J.P. Basto and S.G.S. Alcalá (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024.

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46 (4), 1–37.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21 (9), 1263–1284.

Jinka, P., G.S.V.R.K Rao and K. Sundararaman (2012). ClaimPerfect: An application case study of machine learning techniques towards a fraud management system for warranty claims processing. In *Proc. ICUIMC*. Kuala Lumpur, Malaysia.

Kalbfleisch, J.D., J.F. Lawless and J.A Robinson (1991). Methods for the analysis and prediction of warranty claims. *Technometrics*, 33 (3), 273–285.

Kaufman, S., S. Rosset, C. Perlich and O. Stitelman, (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6 (4), 1–21.

Lauer, M. (2022). The consequences of making goodwill repairs. *Digital Dealer*. https://digitaldealer.com/news/the-consequences-of-making-goodwill-repairs-2/25732/

Park, H.J., T. Kim, Y.S. Kim, J. Min., K.W Sung and S.W. Han (2022). CRFormer: Complementary reliability perspective transformer for automotive components reliability prediction based on claim data. *IEEE Access*, 10.

Rai, B. and N. Singh (2022). Reliability analysis and prediction with warranty data: Issues, strategies and recent advances. *European Journal of Operational Research*, 213 (3), 519–530.

Shan, Q., Y. Hong and W.Q Meeker (2020). Seasonal warranty prediction based on recurrent event data. *The Annals of Applied Statistics*, 14 (2), 929–955.

Vintr, Z. and M. Vintr (2012). Warranty costs assessment for automotive applications. *Advanced Materials Research*, 433–440, 6108–6115.

Wu, S. (2012). Warranty data analysis: A review. *Quality and Reliability Engineering International*, 28 (8), 795–805

Insights of Agriculture 5.0: State of Play and Future Prospects

Elif Uçkan Dağdemir[1]

Abstract

Agriculture sector has experienced many substantial changes accompanied by industrial developments. In 2020s, global challenges like population increase, climate change, water supply dwindles entail a structured and solid transformation of agriculture sector. The conversion of the sector from traditional farming methods to digital farming and agriculture systems has become a necessity for the efficiency and sustainability. In this respect, Agriculture 5.0 is expected to change the rules of the game within the agricultural transformation process. Technological advances of Industry 5.0 like artificial intelligence, internet of things, real time data analytics, and the other technology tools have to be used to serve for bioeconomy and agroecological purposes.

The aim of this paper is to shed a light to the newly-evolving Agriculture 5.0, and evaluate future prospects via elaborating previous industrial and agricultural transformations. In this respect, the first part of the paper covers developments and progress in agriculture sector following the path from Agriculture 1.0 to Agriculture 5.0 that are revealed along with the improvements in industrial sector. At the second part, the scope of Agriculture 5.0 and Industry 5.0 are examined in brief. At the last part, a future perspective is anticipated for Agriculture 5.0, pointing out the main struggles.

Key Words: Agriculture 5.0, Industry 5.0, Bioeconomy, Agroecology.

JEL Codes: O33, Q16, Q18.

Tarım 5.0 İçgörüleri: Mevcut Durum ve Gelecek Perspektifleri

Özet

Tarım sektörü endüstriyel ilerlemelerle birlikte önemli değişimler geçirmektedir. 2020'li yıllarda nüfus artışı, iklim değişikliği, su arzındaki azalış tarım sektöründe yapısal ve somut bir dönüşümü gerekli kılmaktadır. Tarım sektörünün geleneksel çiftçilik metotlarından dijital çiftçilik ve tarım sistemlerine dönüşmesi etkinlik ve sürdürülebilirlik açısından zorunluluk arz etmektedir. Bu bağlamda Tarım 5.0'ın, tarımsal dönüşüm sürecinde oyunun kurallarını değiştirmesi beklenmektedir. Endüstri 5.0'ın kullandığı yapay zekâ, nesnelerin interneti, gerçek zamanlı veri analitiği ve benzeri teknolojik araçlar, biyoekonomi ve agroekolojik amaçlara hizmet edecek şekilde kullanılmalıdır.

Bu çalışmanın amacı, yeni gelişmeye başlayan Tarım 5.0 uygulamalarına ışık tutmak ve önceki endüstriyel ve tarımsal dönüşümler doğrultusunda Tarım 5.0 için gelecek perspektifleri sunmaktır. Bu doğrultuda çalışmanın ilk bölümü, Tarım 1.0'dan Tarım 5.0'a sektörün yaşadığı dönüşümleri, endüstri sektöründeki ilerlemelere paralel şekilde incelemektedir. İkinci bölümde, Tarım 5.0 ve Endüstri 5.0 kısaca tanıtılmaktadır. Son bölümde, Tarım 5.0 için temel zorluklar belirlenerek bir gelecek perspektifi sunulmaktadır.

Anahtar Kelimeler: Tarım 5.0, Endüstri 5.0, Biyoekonomi, Agroekoloji.

JEL Kodları: O33, Q16, Q18.

[1] PhD; Full Time Professor of Economics, Anadolu University Faculty of Economics and Administrative Sciences, Yunus Emre Campus, 26470, Eskişehir, Türkiye
email: euckan@anadolu.edu.tr, ORCID: 0000-0003-1656-0275.

1. FROM AGRICULTURE 1.0 TO AGRICULTURE 5.0: DEVELOPMENTS AND PROGRESS IN AGRICULTURE SECTOR

Agriculture sector has been following the improvements within industrial sector. Agriculture 1.0 which is nominated as the first mark for the developments in agriculture sector began in 1784, following the Industry 1.0, started in 1760. Industry 1.0 was characterized by the steam and water engines, and mechanization of the processes. Coal was the main energy source while railroads were the unique opportunity in transportation during Industry 1.0 era. The significant sectors in industry were textiles and steel. The Industry Revolution which is referred to as Industry 1.0 not only brought large usage of machines within industry, but also induces significant changes in society by transforming it from the agriculture-based society to the industry-based one (Zhang and Yang, 2020, 95). Industry 1.0 may simply be referred to as *mechanization* era.

Agriculture 1.0 which followed the path of Industry 1.0 was characterized by the usage of animal power in agricultural production. Low-skilled labor and low productivity were the other features of the era. In spite of low skills and labor-intensiveness of the agriculture sector, it was successful in feeding the population which mostly employed in agriculture (Kovács and Husti, 2018, 39).

Industry 2.0 was featured with commissioning of internal combustion engine within industry sector in 1870. Electricity, gas and oil were started to be used in industry and mass production was the outcome. Industry 2.0 may simply be referred to as *electrification* era.

Starting in 1950s, Agriculture 2.0 was lagged behind the Industry 2.0 in adapting the improvements. The introduction of internal combustion engine accompanied with the usage of herbicides, fertilizers and other related new inputs started to increase the productivity of the agriculture sector.

Industry 3.0 began in 1960s with the effective implementation of information technologies and automation in production. Computer-based systems replaced the electro-mechanical systems of Industry 2.0. Industry 3.0 may simply be referred to as *information technologies* era.

Agriculture 3.0 started in 1990s, adopting the computer-based systems and introducing the precision agriculture along with the public use of the military GPS (Global Positioning Systems) signals (Zambon et.al., 2019, 4). At the Agriculture 3.0 era, telematics that are widely used in transportation started to be implemented in agriculture, enabling to control the supply chain processes. The term of *precision agriculture* was started to be used during this era.

Industry 4.0 initiated by Germany with the expectation of being the leader of industrial manufacturing all around the globe, and sustaining its hegemony in the equipment manufacturing. German government announced the launch of Industry 4.0 during the Hannover Expo in 2011. Industry 4.0 gives priority to machine learning, internet of things and big data analytics. Undoubtedly, the revolutions within the scope of Industry 4.0 improve not only the industrial manufacturing, but also the processes of marketing, international trade, finance and the other related components of a comprehensive firm management (Schwab, 2017). Industry 4.0 may simply be referred to as *intelligence* era.

Main feature of Industry 4.0 is to build a *smart* industry by interconnecting machines and related devices. In this respect, the main motive of Industry 4.0 is to initiate and manage entire automation, reducing human element within production processes via more integrated machine learning based intelligence. However, the main of aim of Industry 4.0 is not different from the previous industry revolutions: achieving the mass production by using new information technologies (IT).

Industry 4.0 has led to some controversies. The first controversy is the burden of high costs of adapting the required IT, especially for the small and medium sized enterprises (Hirsch-Kreinsen, et.al. 2016). The second controversy is the possibility of negative employment effects and social consequences which may result from the replacement of humans by robots (Hirsch-Kreinsen, 2016, 4-7). Another controversy

is the high demand for skilled labor, capable of working along with IT. Possibly the most important feature of Industry 4.0 is its dependency on linear economy. These facts have led to an increased negative perception of Industry 4.0, and paved the way to the rapid developments referred to as Industry 5.0.

As in the previous stages, Agriculture 4.0 has followed the path of Industry 4.0. Agriculture 4.0 initiated the usage of digital data in all agricultural processes. Nevertheless, due to significant differences between industrial manufacturing and agriculture sector both in production and supply chain management, the adoption of digital revolutions within agriculture sector has required a delicate endeavor.

In the era of Agriculture 4.0, discrete features of agricultural production have become more apparent. Agricultural production consisting of crop and animal production, generates a sustainable production and nutrient cycle since crop production incorporates cultivation of plants for the nutrition of human beings, animal feeding, as well as for generating energy in the form of biomass and biofuels. The inter-connection of real and virtual world within the agricultural production is a challenge for agriculture sector since precision production has been possible along with monitoring and feeding animals, more efficiency in management of arable land and processes (Braun et.al. 2018, 979-980). Intelligent planning, monitoring, controlling and documentation are the components of Agriculture 4.0 as of Industry 4.0. Agriculture 4.0 is usually referred to as *smart farming* or *digital agriculture*.

2. INDUSTRY 5.0 AND AGRICULTURE 5.0: A CONCEPTUAL OVERVIEW

Industry 5.0 may be conceptualized as a tailor-made revolutionary process which delicately encompasses sustainability, energy and environmental issues that are lagging behind at the Industry 4.0 era. It has to be mentioned beforehand that Industry 5.0 of 2020s is aiming to bridge the innovations of human experts and professionals with the intelligent machinery. In other words, Industry 5.0 would like to reintroduce human being into manufacturing. The expectations from Industry 5.0 are to integrate technology with the human talent and expertize. Personalization within production is another improvement of Industry 5.0 which enables consumers to choose customized commodities produced upon their preferences and special demands. Unlike Industry 4.0 which depends on mass production, Industry 5.0 gives priority to consumer satisfaction. Industry 5.0 changes the meaning and tasks of robots, introducing the next generation robots, referred to as *cobots*. A cobot is a collaborative robot that works next to human worker without replacing him and can learn from its real counterpart (Nahavandi, 2019, 3).

Demir et.al. (2019) emphasizes two visions of Industry 5.0. The first vision referred to as *human-robot co-working*. According to this vision, humans will take the responsibility of creative tasks while robots will do the remaining. The second vision of Industry 5.0 is *bioeconomy*. *Bioeconomy* vision postulates that smart utilization of biological resources within industry sector will enhance the balance among industry, ecology and economy (Demir et.al. 2019, 690). According to the precise and clear definition of the European Commission, bioeconomy incorporates the production of biomass, its transition into food, materials, products, and generation of bioenergy (European Commission).

Circular economy replaces linear economy within the framework of Industry 5.0. Industry 5.0 is human-centric, replacing the focus from shareholder to stakeholder value (Nahavandi, 2019, 11). Sustainability and resilience are the core elements of Industry 5.0 (Atif, 2023, 2158-2159). It prioritizes human-machine collaboration and sustainability, and aims to incorporate human creativity and ethical aspects with technological advances.

As mentioned before, Agriculture 4.0 follows the path of Industry 4.0 on the implementation of automation and digitalization with the aim of increasing productivity and optimization in agricultural production. Industry 4.0 which encloses real time data analytics and related IT has eased smart agriculture and precision farming, sidelining the human factor. In conjunction with Industry 5.0, Agriculture 5.0 is aiming to integrate latest technological advances with larger societal and

environmental aims.

3. A PROMISING FUTURE PROSPECT FOR AGRICULTURE 5.0?

Despite the abovementioned challenging tasks, is it possible to advocate a promising future prospect for Agriculture 5.0? The answer is yes. How? By a strong and structured collaboration between industry and agriculture within the framework of bioeconomy. Another promising future prospect is to incorporate the agroecology principles into the bioeconomy framework.

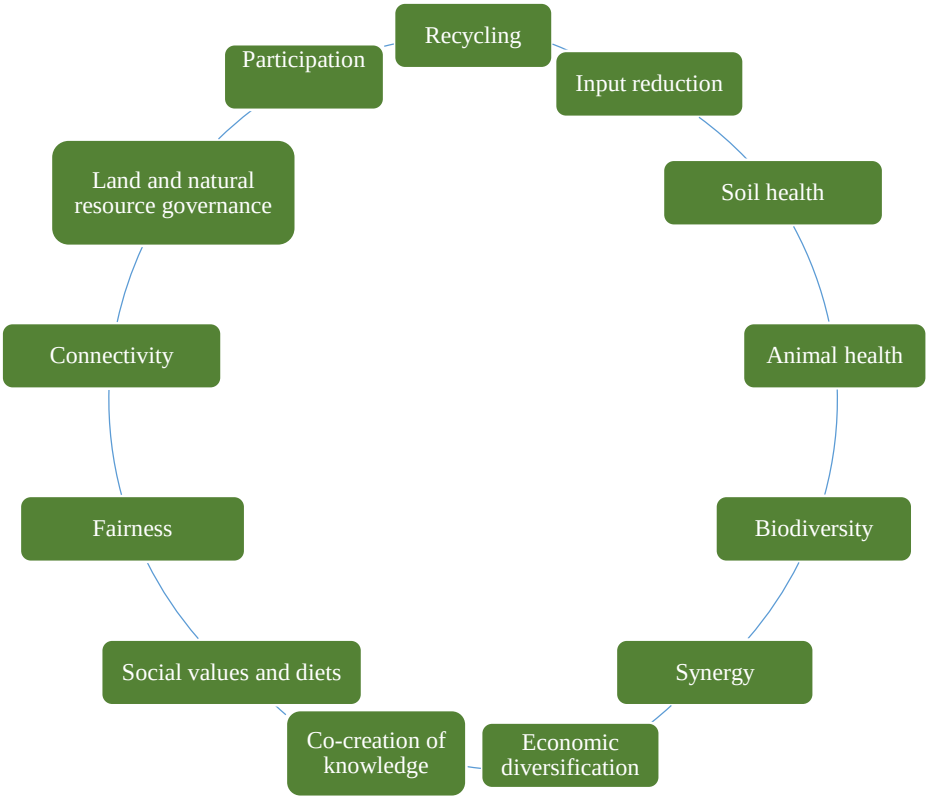


Figure 1. Thirteen Agroecological Principles
Source: Figure is prepared by author based on Wezel, et.al: 2020: 7.

Wezel et.al. (2020, 7) present the agroecological principles under thirteen headings. These principles are given at Figure 3.1. According to Wezel et.al. (2020, 7), recycling principle encompasses the usage of local renewable resources. Input reduction principle aims to increase local input usage. Soil health principle propose enhancing soil health. Animal health principle is quite similar to the soil health one. Biodiversity principle has the goal of maintaining and enhancing diversity. Synergy principle aims to enhance ecological interaction amongst the elements of agroecosystems. Economic diversification principle advocates to diversify the small-scale farmers, providing them financial facilities for more independence and respond to consumer demand. Co-creation of knowledge principle deals with sharing and spreading knowledge within agroecological eco-systems. Social values and diets principle aims to develop food systems based on culture, identity, tradition of locals in order to sustain healthy, diversified, seasonally and culturally adequate diets. Fairness principle proposes to establish a fair food system for all participants, especially small-scale producers. Connectivity principle is about ensuring proximity and confidence between producers and consumers by re-embedding food systems into local communities and economies. Land and natural resource governance principle has the goal of strengthening institutional framework so as to incorporate locals in maintaining natural and genetic resources. Participation principle proposes to provide decentralized governance of agriculture and food systems via adaptive local management.

It is evident from the principles that agroecological agriculture places local farming communities at the center of the food system. Decentralized decision making procedures, financial facilities and assistance for small-scale farmers are the core elements. Respecting for nature and animal wealth are among the main values. Sustainability constitutes the main body of the proposed system. Consequently, the success of Agriculture 5.0 will be contingent upon a proper blending of productivity and optimization purposes with societal, environmental and humanitarian prospects. Here the challenging task for Agriculture 5.0 is to embed the technological advances into the agroecological agriculture in order to remedy the negative social and ecological consequences of the previous agricultural innovations that were mostly served for *industrial agriculture*.

Agroecological agriculture in which the technological advances of industry are embedded will significantly serve for the proper functioning of bioeconomy. In such a framework, artificial intelligence, internet of things, real time data analytics and all the related technological improvements and facilities will be integral parts of the system as in the previous revolutionary times, but with a significant difference: In this case, they will be the instruments of achieving sustainability, strengthening bioeconomy and prioritizing societal purposes via integrating local agricultural stakeholders within the entire agricultural system.

4. CONCLUSION

Agriculture sector is undergoing a profound and promising transformation with Agriculture 5.0. It is being re-shaped along with the technological improvements of Industry 5.0. This transformation is generated within the framework of bioeconomy. The efficiency of Agriculture 5.0 will exponentially be increased by the incorporation of agroecological principles within the bioeconomy framework. Artificial intelligence, internet of things, real time data analytics and all the related technological advances have to be main capital of the agriculture sector. In such a system, the main motive will be sustainability, societal and environmental purposes with and inclusive manner. Agriculture 5.0 can be referred to as *inclusive agriculture*, incorporating local communities and stakeholders into the agriculture system, giving more floor to local producers and consumers demand. A well-structured decentralized decision-making system in agriculture within Agriculture 5.0 may pave the way to a more resilient sector.

Certainly, such a transformative shift encloses some worth-mentioning struggles and a big question mark. The struggles can be classified into three main groups. The first group incorporates economic struggles. High costs due to high technology and skilled labor have to be met. When small-scale farmers are considered, these high costs are not easy to overcome. The second group consists of technological struggles. Adequate infrastructure for technology has to be provided, and labor force has to have the required skill to adapt and implement the related technology. Third group of struggles encompasses political will and support. There has to be a firm and sustainable political commitment to overcome these economic and technological struggles and to enhance the effectiveness of improvements. With a solid political commitment, countries may give financial, technical and educational support for the agricultural transformation under Agriculture 5.0.

Then the question mark comes: Can Agriculture 5.0 with its drivers of bioeconomy and agroecology meet the rising global food demand? The answer to this question will be addressed in a subsequent study conducted by the author.

REFERENCES

Atif, S. (2023). Analyzing the Alignment between Circular Economy and Industry 4.0 Nexus with Industry 5.0 Era: An Integrative Systematic Literature Review. *Sustainable Development*, 31, 2155-2175.

Braun, A.T., E. Colangelo and T. Steckel (2018). Farming in the Era of Industrie 4.0. *Procedia CIRP 72*, 979-984.

Demir, K. A., G. Döven and B. Sezen (2019). Industry 5.0 and Human-Robot Co-working. *Procedia Computer Science*, 158, 688-695.

European Commission. https://environment.ec.europa.eu/strategy/bioeconomy-strategy_en (accessed May 5, 2026).

Hirsch-Kreinsen, H. (2016). Digitalization of Industrial Work: Development Paths and Prospects. *Journal of Labour Market Research*, 49 (1), 1-14.

Hirsch-Kreinsen, H., J. Weyer and J. D. M. Wilkesmann (2016). "Industry 4.0" as Promising Technology: Emergence, Semantics and Ambivalent Character.

Kovács, I. and I. Husti (2018). The Role of Digitalization in the Agriculture 4.0 – How to Connect the Industry 4.0 to Agriculture? *Hungarian Agricultural Engineering*, 33, 38-42.

Nahavandi, S. (2019). Industry 5.0 – A Human-Centric Solution. *Sustainability*, 11, 4371, 1-13.

Schwab, K. (2017). *The Fourth Industrial Revolution*. Crown Publishing Group: New York.

Wezel, A., B. G. Herren, R. B Kerr., E. Barrios., A. L. R. Gonçalves and F. Sinclair (2020). Agroecological Principles and Elements and Their Implications for Transitioning to Sustainable Food Systems. A Review. *Agronomy for Sustainable Development*, 40 (40), 1-13.

Zambon, I., M. Cecchini., G. Egidi, M. G. Saporito and A. Colantoni. (2019). Revolution 4.0: Industry vs. Agriculture in a Future Development for SMEs. *Processes*, 7 (36), 1-16.

Zhang, C. and J. Yang (2020). *A History of Mechanical Engineering*. Springer Nature Singapore, 95-135.

Uncovering Turkish Audit Firms' Transparency Report Textual Attributes Through Computer Based Approaches and Linking Them to Audit Quality

Sibel Dinç Aydemir[1], Yusuf Sinan Akgül[2], Barış Özcan[3] and Mine Aksu[4]

Abstract

This paper assesses the textual attributes of transparency reports (TRs) prepared by audit firms (AFs) in Türkiye between 2013-2020. To achieve this goal, we utilize a range of computer-based techniques and expertise, including natural language processing, named entity recognition, business text processing, web scraping, data mining, and bot building. The findings reveal that TRs exhibit boilerplate tendencies, lacking full transparency, are characterized as medium-to-hard documents with a generally positive tone and diverse EoBP at the firm level, pointing to significant barriers in achieving sustainable reporting practices.

This paper also investigates AF-based determinants of TR textual attributes. Regression model results indicate the impact of some AF characteristics (e.g., license duration, number of partners, the Big 4, female ratio, non-audit services) on most textual attributes.

Finally, we explore the role of textual features, alongside AF and audit partner (AP) attributes, in influencing audit quality. To address potential endogeneity concerns, the study adopts an instrumental variables probit model. The analysis reveals that the readability of TRs by Fog index significantly influences audit quality. Factors such as AF tenure, AP busyness, CF's sales ratio, and loss contribute to audit quality, aligning with prior research.

Keywords: Audit Quality; Transparency Reports; Readability; Natural Language Processing; Business Text Processing.

JEL Codes: C26, C88, G18, M42.

Bilgisayar Tabanlı Yaklaşımlarla Türk Denetim Kuruluşlarının Şeffaflık Raporu Metin Özelliklerinin Ortaya Çıkarılması ve Denetim Kalitesiyle İlişkilendirilmesi

Özet

Bu çalışma, Türkiye'de 2013–2020 yılları arasında denetim firmaları (DF'ler) tarafından hazırlanan şeffaflık raporlarının (ŞR'ler) metinsel özelliklerini incelemektedir. Bu amaca ulaşmak için doğal dil işleme, isimlendirilmiş varlık tanıma, iş metni işleme, web kazıma, veri madenciliği ve bot geliştirme gibi bilgisayar tabanlı çeşitli teknikler ve uzmanlıklardan yararlanılmıştır. Bulgular, ŞR'lerin tam şeffaflık göstermediğini şablon formlar olma eğilimi sergilediğini ortaya koymaktadır. Raporlar, ortadan zora bir okunabilirlik düzeyine sahip belgeler olarak ortaya çıkmakta, genel olarak olumlu bir tona sahip olmakta ve firma düzeyinde çeşitlilik gösteren bir temel prensip ve iyi uygulamalara vurgu içermektedir. Ayrıca, Fog endeksinden türetilmiş, Türkçeye uyarlanmış özgün bir okunabilirlik ölçeği geliştirilmiştir. Çalışma ayrıca, denetim firmalarına özgü belirleyicilerin şeffaflık raporlarının metinsel özellikleri üzerindeki etkilerini incelemektedir. Regresyon modeli sonuçları, lisans süresi, ortak sayısı, dört büyük denetim firmasından olma durumu, kadın ortak oranı ve denetim dışı hizmetlerin varlığı gibi bazı denetim firması özelliklerinin, metinsel özelliklerin çoğu üzerinde anlamlı etkileri olduğunu göstermektedir. Son olarak, metinsel özelliklerin, denetim firması (DF) ve denetim ortağı (DO) nitelikleriyle

[1] Sibel Dinç-Aydemir, Assoc. Prof, Management Information Systems, Fatih Sultan Mehmet Vakif University, Halic Campus, 34421, Beyoğlu / İstanbul, Turkiye
email: sdaydemir@fsm.edu.tr), ORCID: 0000-0003-1897-0913.
[2] Yusuf Sinan Akgül, Professor of Computer Engineering, Gebze Technical University, 41400, Gebze Kocaeli, Türkiye
email: akgul@gtu.edu.tr), ORCID: 0000-0001-8501-4812.
[3] Barış Özcan, ForInvest Yazılım ve Teknoloji Hizmetleri A.Ş., İstanbul, Türkiye,
email: barisozcan963@gmail.com, ORCID: 0000-0001-7555-9122.
[4] Mine Aksu, İstanbul, Türkiye
email: maksu@sabanciuniv.edu, ORCID: 0000-0003-3861-6823.
The authors gratefully acknowledge the helpful and substantive comments of DATAMACLEA'26 on concepts related to this paper.

birlikte denetim kalitesi üzerindeki rolü araştırılmaktadır. Olası içsellik sorunlarını gidermek amacıyla çalışmada araç değişkenli probit modeli kullanılmaktadır. Analiz sonuçları, Fog endeksi ile ölçülen şeffaflık raporlarının okunabilirliğinin denetim kalitesi üzerinde anlamlı bir etkisi olduğunu ortaya koymaktadır. Ayrıca, DF'nin ardışık görev süresi, DO'nun iş yoğunluğu, müşteri firmanın satış oranı ve zarar durumunda olup olmadığı gibi faktörlerin denetim kalitesini etkilediği ve bu bulguların önceki literatürle uyumlu olduğu görülmektedir.

Anahtar Kelimeler: Denetim Kalitesi, Şeffaflık Raporları, Okunabilirlik, Doğal Dil İşleme, İş Metni İşleme, Varlık İsmi Tanıma, Belge Benzerliği.

JEL Kodları: C26, C88, G18, M42.

1. INTRODUCTION

Due to the efforts of predominantly "whistleblower" employees, the business world witnessed unprecedented accounting-auditing scandals and fraudulent financial reporting in the early 2000s. Consequently, the entire world has undergone significant and transformative changes in terms of legal regulations, new standards, and institutional arrangements aimed at rebuilding trust in the audit profession, hence at restoring the reputation and credibility of financial reporting systems. These regulations have also prompted AFs to be more transparent with all their stakeholders, particularly their client firms and investors, regarding their firms' audit activities and internal governance practices. This was also influenced by calls for transparency in the governance of AFs, previously known as "black boxes" (Wyman, 2004; Huddart, 2013). Prompted by the European Union (EU) Directive, mandatory TR disclosure has subsequently followed these resounding calls and stringent reforms, ultimately capturing the attention of numerous researchers. In fact, considerable efforts have been exerted by scholars focusing on audit venues in general and on TRs in particular. Through this valuable literature, we have gained insights into various aspects: (1) whether these practices by AFs in different countries align with jurisdictional TR legislation or have improved since the first promulgation year (Cular, 2009; Pivac and Cular, 2012; Cular, 2014; Fu et al., 2015), (2) whether the Big 4 in certain countries meet the minimum information requirements in TRs (Pheijffer, 2010), (3) the disclosure levels of specific sections within them (Deumes et al., 2012), and the extent of corporate governance disclosure in these reports (La Rosa et al., 2019). Additionally, studies have explored whether disclosure or corporate governance levels in TRs contribute to improved investor confidence (La Rosa et al., 2019). Unfortunately, this body of research often focuses on either one to two years of TRs (Fu et al., 2015), specific AFs (Pheijffer, 2010), or certain sections of these reports (Deumes et al., 2012).

However, several questions persist without clear answers. Li (2010) highlights the drawbacks of manual analysis and advocates for the use of computer-based techniques in textual analyses to extract more information or attributes. The research on TRs discussed earlier relies on interpretive (manual) analysis. While recent studies have explored the readability of various documents, including annual reports (Bodnaruk et al., 2015; Li, 2019; El-Haj et al., 2020), corporate disclosures (Siano and Wysocki, 2018), and key audit matters (Smith, 2021; Seebeck and Kaya, 2023), the extent to which AFs disclose TRs that are comprehensible and readable for their intended audience remains unknown. Despite being rich sources of information that inherently contain indicators of audit quality, TRs have not been thoroughly examined in this regard. While numerous studies have highlighted textual attributes in the context of corporate or financial disclosures, including tone (Loughran and McDonald, 2011; Ertugrul et al., 2017; Ehrmann and Talmi, 2020; Teng and Han, 2022; Demers et al., 2022), there is a notable gap in our understanding concerning the tone of transparency reports by audit firms known for their seriousness and confidentiality. Moreover, certain studies have presented evidence of a boilerplate effect on documents like 10-K reports (Dyer et al., 2017) and human resource disclosures (Demers et al., 2022). Yet, the inquiry persists regarding whether transparency reports are indeed boilerplate documents that lack distinctive and unique information sufficient for their users.

Moreover, a body of research has raised concerns regarding the content and utility of mandatory reports envisaged by standard setters to enhance transparency and mitigate information asymmetry (Gray et al., 2011). Furthermore, there is scrutiny regarding the emphasis on using clear, concrete, and understandable language in these reports (Seebeck and Kaya, 2023). Additionally, a recent event

occurred when the Financial Times reported the collapse of Silicon Valley Bank and Signature Bank, both audited by a Big 4 accounting firm, within days of disclosing their annual reports. This development prompts a reevaluation of whether the regulatory shifts aimed at increasing transparency and reducing information asymmetry globally, instituted nearly twenty years ago, have indeed achieved the objectives envisioned by standard setters in the 2000s. As researchers, it appears that we have not thoroughly expanded our exploration of audit quality and illuminated audit process comprehensively.

This paper introduces several contributions to the accounting and audit research. First, to the best of our knowledge, the computer-based approach, the advantages of which are emphasized by Li (2010) will be employed on TRs for the first time. Second, textual features such as readability, tone, similarity which have been largely argued on corporate disclosures such as annual reports, 10-K files, will be investigated on transparency reports for the first time. Third, unlike prior research which overwhelmingly focused on (1) one to two years, (2) specified sections of TRs, and (3) specific types of AFs (i.e., Big 4) due to manual content analysis in their design, this research includes all TRs disclosed by all AFs which conducted audits on capital markets in Türkiye between 2013-2020.

Fourth, some research (Deumes et al., 2012) focused partially on the link between disclosure level and audit quality at country level, hence provided insignificant evidence for this link. Unlike this research, we will be able to observe the same link directly and entirely in one country sample by removing potential differences between countries. Türkiye is a peculiar and distinct research setting for two reasons. Firstly, Public Oversight Accounting and Auditing Standards Authority (POAASA) and Capital Markets Board (CMB) of Türkiye have already disclosed the current data regarding audit realm to the public, however the data is unfortunately not provided for the potential users including researchers on a longitudinal basis. As such, TRs are also rich sources containing audit firm and partner-based attributes on a yearly basis, enabling a longitudinal data analysis. This study conducted a thorough and comprehensive data collection by manually or computer-based techniques such as web scraping, bot building, data mining from CMB, Public Disclosure Platform (PDP) web sites, and TR documents. This is quite rare for the auditing research in Türkiye. Secondly, the criteria for independent audits have recently changed, effective November 30, 2022. As a result, firms' financials from 2021 and 2022 will determine which firms will undergo audits in 2023. This new regulation will significantly alter the profile and scale of client firms audited by audit firms (AFs). Our research will serve as a crucial reference point for future studies, as our sample includes client firms identified under the previous regulations.

Finally, alongside the contributions outlined above, we expect the most significant contribution of these reports to be in the monitoring, regulatory, and policy making activities of the authorities. These reports stand out due to their dense textual content and lengthy document formats. These reports play a critical role in the proper and reliable functioning of financial markets since they reveal whether audit firms possess strong audit teams, how they build robust corporate governance mechanisms, their adherence to independence principles, how they mitigate risks in audit processes, and the strength of their financial structures. In this regard, the research findings will provide valuable insights for regulators, ranging from the assessment of whether the objectives pursued through these reports are achieved, to the development of more functional policies. This will also serve as a functional foundation for establishing a stronger base for sustainable reporting practices. These reports, which are the result of significant time and effort on the part of their preparers, could suffer from ineffectiveness and a lack of sustainability if not thoroughly and deeply analyzed. However, when analyzed using computer-based or AI-driven methods, these reports have the potential to transform into a powerful, sophisticated, and skillful monitoring tool.

The study also embodies advanced methodological strengths. Firstly, the research design of the study encompasses rich computer-based techniques combining rule-based approaches (e.g., named entity recognition, dictionary approach) with statistical approaches (e.g., Turkish scale development, Fog index, EoBP construct). Many variables in the research model have been attained through computerized processes. Previous studies on TRs overwhelmingly employed manual (interpretive) content analysis.

Therefore, most of these studies base on one to two years since these manual methods take substantial time and effort.

Secondly, in this study, we developed a new customized Fog scale appropriate for Turkish language. While criticism regarding the complex word threshold value persists, the Fog index, along with its parameters, still stands as a simple, useful, and generic scale for computing the readability of a text. Yet, for example, Bog index is utterly peculiar to English in terms of its components such as sentence bog, word bog and pep. It is also embedded on StyleWriter, not an open-source program, hence it is not easily adaptable to other languages. Consequently, the studies using Bog index are still limited (Bonsall et al., 2017; Kuang et al., 2020; Rjiba et al., 2021; Blanco et al., 2021). Thus, researchers may use our scale development approach to tailor Fog index parameters to their languages other than English without needing any commercial software.

Thirdly, in measuring document similarity in TRs, we utilize WCopyFind, a unique application within plagiarism detection systems. Unlike the MOSS software used by Li (2019), which could not disclose its computational method due to intellectual property restrictions, WCopyFind is an open-source software developed by the University of Virginia in 2004. This offers a significant advantage to future researchers for follow-up studies. Furthermore, document similarity research often relies on cosine similarity, which has drawbacks compared to our use of WCopyFind. Cosine similarity requires identifying unique dimensions and transforming documents into vectors before conducting file comparisons. It also considers the direction or orientation of vectors. In contrast, WCopyFind directly compares text files and computes similarity based on size or magnitude, aligning with the objectives of our study. Moreover, WCopyFind addresses differences in document length without requiring any scaling process.

Fourth, the research design of the studies in auditing research mainly includes audit firm and client firm-based attributes. However, our study includes audit partner attributes, reports' textual attributes into the research model. This design makes our study stronger for two reasons. Firstly, some studies call for the research design including audit partner attributes since there are still questions without clear answers (please see Lennox and Wu, 2018 for a detailed review). Secondly, this study deals with endogeneity problem emphasized greatly on the literature, through suggesting an instrumental variables estimation model (IV) which is still rare in the auditing research.

By parsing all TRs through computer-based techniques we first find that TRs are medium-to-hard documents. EoBP differentiates across audit firms, referring to its potential as a distinctive textual feature. Further, reports have a relatively positive tone. The similarity results indicate that AFs use remarkably similar content in their reports across the years, suggesting a boilerplate effect. Lastly, the transparency scores of the reports suggest that AFs are still not fully transparent, despite incremental changes over the years.

The analysis examining the impact of AF characteristics on textual attributes produced interesting findings. Factors such as license duration, number of partners, Big 4 affiliation, ownership structure, provision of non-audit services, report timeliness, number-to-letter ratio, and market share significantly influence scores on readability scales. Additionally, foreign licensure increases readability difficulty according to the Turkish Fog scale, while the presence of female partners improves readability only in the standard Fog model. Yearly effects on readability were not statistically significant. License duration, Big 4 affiliation, foreign licensure, sanctions, and non-audit services positively influence EoBP, whereas female partner ratio, number-to-letter ratio, and market share have a negative impact. The results highlight AFs' efforts in managing their image through TRs, particularly concerning sanctions and market share. EoBP shows a significant increase during the years 2016-2020.

The intra-similarity of TRs (boilerplate effect) is influenced by license duration, Big 4 affiliation, ownership structure, and the number-to-letter ratio. Specifically, license duration, Big 4 affiliation, and the number-to-letter ratio decrease the boilerplate tendency in these reports, while ownership structure

tends to increase it. Significant positive year effects are observed for the years 2015-2020. None of these features significantly influence the transparency score of reports. Notably, there is an increase in the transparency level of reports for the years 2016-2020. Higher report timeliness is associated with increased readability and reduced EoBP. Additionally, TRs disclosed earlier than the legal deadline tend to have a more positive tone.

Unlike prior research, while linking all textual attributes to audit quality, we proposed treating textual attributes as endogenous variables based on our findings and earlier outputs of our research project (Ozcan et al., 2023; Aydemir et al, 2022; Aydemir et al., 2021) and estimated an instrumental variables probit model. Our findings indicate that AFs producing more readable TRs tend to issue more modified audit opinions, suggesting higher audit quality. Additionally, audit firm tenure, partner busyness, and client firm sales show negative associations with audit quality, consistent with previous research. A modified audit opinion is more likely in years when the client firm has experienced a loss. Yearly effects are significant across all years, indicating a decreasing trend in modified audit opinions over time. While other textual attributes such as tone, EoBP, document similarity, and transparency level did not yield notable effects, this study provides evidence for the significant relationship between a specific textual attribute (readability) and audit quality for the first time.

2. CONCEPTUAL FRAMEWORK

2.1. Institutional Background

Due to the fradulent financial reporting practices and auditing scandals, a great number of regulations have been released and entered into force, most of which are strict indeed. For example, Sarbanes-Oxley Act enacted by the United States Congress in 2002, firstly supports audit and public disclosure (Cornell Law School, 2023). It also sets out corporate responsibility, increased transparency, increased punishments for those lawbreakers, accounting regulation and new protections through reforms and extentions. Then, EU is considered as the pioneer of important regulations with Article 40 -TR Section of the 8th Directive (Official Journal of the European Union, 2006). As of June 2008, this article mandates the issuance of TRs in member countries. Following Article 40, member states began to disclose their reports at different times. For example, Spanish AFs started their reporting practices following the Spanish Audit Act 12/2010 (Zorio-Grima et al., 2018). In Croatia, these practices launched in 2011, after adopting the Code of Ethics for Accountants by the Croatian Audit Assembly in 2010 (Cular, 2014). Non-member countries also adopted Article 40 of the EU, and in this direction, Australia implemented compulsory TR disclosure by AFs in 2013 (Fu et al., 2015).

The initial regulation on mandatory TRs in Türkiye was introduced through Article 36 of the Independent Audit By-Law. It was enacted by the Public Oversight Accounting and Auditing Standards Authority (POAASA) on December 26, 2012. This article regulates matters pertaining to the timing, responsible entities, publication locations, and mandatory components to be included in TRs.

2.2. Literature Review

Transparency reporting has emerged as a relatively recent development for AFs, drawing increased attention from standard setters, regulators, and researchers interested in its potential impact on market participants' decisions and perceptions. The initial scholarly papers on this subject began to surface shortly after 2008, coinciding with the European Union's inaugural mandate requiring AFs to publicly disclose information about their ownership, management, internal governance mechanisms, and audit processes in these reports.

These studies were descriptive and qualitative content analysis papers that examined the contents of TRs. They often involved a manual search to assess compliance with the disclosure requirements mandated by regulators in their respective jurisdictions (Cular, 2009; Pivac and Cular, 2012; Cular, 2014). An illustrative instance is the work of Fu et al. (2015), which focused on the 2013 inaugural TRs

of 21 Australian AFs. In this study, two co-authors conducted independent manual coding of TRs against 11 categories of disclosure requirements. The findings revealed that while AFs generally met the minimum disclosure standards, variations existed in their governance practices. As another notable contribution, Zorio-Grima et al. (2018) concentrated on TRs disclosed by Spanish AFs and found that (1) there is an increase in mandatory information from 2011 to 2013, (2) large firms and those less dependent on non-audit services (NAS) have higher transparency and disclosure level.

Following Türkiye's early adoption of mandatory transparency reports in 2013, several studies examined Turkish TRs from different perspectives, including content analyses, disclosure practices, and relationships among disclosed information items (Erdogan and Solak, 2016; Tanc and Gumrah, 2016; Erdogan and Kutay, 2016). Other studies focused on determinants of audit firms' revenues and ownership structure, reporting a positive association between female ownership and revenues (Bozcuk, 2018; Ocak, 2021). Cross-country comparisons of Big-4 firms also revealed differences in disclosure quality and reporting practices across jurisdictions (Gurol and Tuysuzoglu, 2016).

Several papers have attempted to construct a TR disclosure index, facilitating the ranking of AFs based on their disclosure quality. Ehlinger (2010) devised a disclosure index after examining 125 reports published in Austria, Germany, and the Netherlands. The analysis revealed variations in TR disclosures among Big 4 and foreign license owners and across different countries and sample periods (i.e., 2007 and 2008). Additionally, it provided evidence for a positive relationship between audit firm size and auditor independence. Similarly, Pivac and Cular (2012), along with Cular (2014), developed a disclosure quality index for 37 TRs on 11 information components mandated by Croatian law. The importance of each information item was determined through a survey conducted on audit experts. Their findings indicated that only 7 AFs had transparent reports, with an improvement in the transparency from 2011 to 2012. Notably, they found no correlation between AFs' total income and their disclosure quality.

Overall, research on TRs primarily uses normative approaches and manual content analysis to assess how well these reports align with legislative requirements regarding the content and extent of information they provide about AFs. Due to the time and effort-intensive nature of manual content analysis, studies typically focus on TRs spanning (1) one to two years, (2) specific firms like the Big 4, or (3) particular sections such as independence compliance and quality control mechanisms. In addition to interpretive content analysis, some studies use interviews and surveys (Pott et al., 2008; Girdhar and Jeppesen, 2018).

As the final strand of literature, closely aligned with the focus of our study, there are comprehensive studies that not only developed disclosure index but also explored the determinants or consequences of these disclosures. For instance, Deumes et al. (2012) manually constructed a TR disclosure index and examined its relationship with audit quality. Analyzing 103 TRs published in the United Kingdom, Austria, and the Netherlands, they calculated a score, but they concentrated on only specific sections such as continuous training policy, independence principles, and quality control systems. Their approach involved a structural equation model comprising 87 questions, with allocations of 53, 21, and 13 questions respectively across these sections. In this model, firm size was the sole audit firm attribute and audit quality was measured by audit opinion, audit fees, and discretionary accruals. In another comprehensive study covering 219 TRs from Germany, the United Kingdom, France, the Netherlands, Belgium, and Austria in 2008, Van Buuren and Koch (2014) developed a scale consisting of 92 questions, manually administered by the authors and three research assistants. Their findings indicated that AFs generally provided broad information on ex-ante control structures rather than detailed disclosures on ex-post quality performance. The study identified audit firm size as a key determinant influencing transparency scores. Furthermore, it highlighted that medium-sized AFs frequently emphasized their global networks and training initiatives, whereas the Big 4 focused prominently on their quality control systems.

Although these studies serve as motivation for our research, our study differs in many respects from them. First, like Van Buuren and Koch (2014), we determined 88 questions aimed to measure transparency score of these reports. Differently, we use mostly computerized techniques to come up with a systematic, objective, descriptive evidence related to transparency scoring. Second, in addition to transparency score, we also aimed to measure other textual attributes such as readability, tone, emphasis on basic principles, document similarity. Deumes et al. (2012) employed interpretive (manual) content analysis, focused on only three parts of transprency reports aimed at calculating a disclosure score and linked only this disclosure score and audit firm size to audit quality. However, more differently, we have incorporated a much more extensive list of audit firm-based attributes alongside those of audit partners and client firms. Furthermore, our study explores the link between a diverse range of textual attributes and audit quality.

On one hand, Li (2010) underscores the limitations of manual content analysis and advocates for computer-based techniques due to their inherent advantages, including improved generalizability, enhanced empirical insights, and ease in subsequent research through non-subjective coding and interpretation. Moreover, these methods enable more powerful empirical tests with larger sample sizes. This perspective has significantly influenced our study. Notably, a substantial body of research has employed computer-based methods to analyze corporate disclosures such as annual reports, corporate disclosures, and key audit matters (Demers et al., 2022; Seebeck and Kaya, 2023; Blanco et al., 2021; Bodnaruk et al., 2021; El-Haj et al., 2020; Smith, 2021; Siano and Wysocki, 2018; Li, 2019). These studies often explore textual attributes such as readability, complexity, clarity, tone, and content similarity. Nevertheless, despite the extensive use of these advanced techniques in financial context, TRs have not yet been analyzed using such methods. This research gap further motivates our study.

On the other hand, some studies have also raised concerns about the growing use of boilerplate disclosures in TRs and corporate reporting, arguing that generic and non-distinctive information may weaken transparency and obscure links with audit quality (Deumes et al., 2012; Dyer et al., 2017; Girdhar and Jeppesen, 2018). These concerns provide further motivation for our research.

2.3. Theoretical Underpinnings and Hypothesis Development

This research draws on agency theory (Jensen and Meckling, 1976), signaling theory (Spence, 1973), stakeholder theories, the incomplete revelation hypothesis, image management. Agency theory posits that conflicts of interest between audited firms and stakeholders result in higher monitoring costs due to information asymmetry. Independent audits mitigate these costs by reducing asymmetry and ensuring trustworthy financial operations. Similarly, transparency reports released by audit firms aim to reduce agency costs for stakeholders by demonstrating meticulous audit practices and processes. Signaling theory suggests that market agents interpret various signals to mitigate information asymmetry. TRs serve as signals to users about audit processes, corporate and financial structures, thereby signaling audit quality.

In contrast to the efficient market hypothesis (Fama, 1970), which suggests that market prices incorporate all available information, some studies argue that extracting useful insights from public data requires significant effort from investors. Consequently, reports that are difficult to read may increase information extraction and processing costs, hindering users from deriving useful information. Additionally, according to the management obfuscation hypothesis (Lang and Lundholm, 1993; Li, 2008), managers may use corporate reports to downplay negative news and emphasize positive aspects, influencing how audit firms present information favorably. Furthermore, guided by stakeholder theory (Freeman, 1983), firms use disclosures to demonstrate their commitment to stakeholders and their interests. Research also indicates that image management drives firms to tailor their language and content in reports to create a favorable impression on stakeholders (Cho et al., 2010, 2012). In summary, there is debate over whether these corporate communication mediums enhance transparency or merely reflect efforts to manage corporate image (Arena et al., 2015).

Previous research indicates that audit firm characteristics significantly impact the quality of disclosures in transparency reports. Petersen and Zwirner (2009) found a positive correlation between transparency and audit firm size. Similarly, some studies demonstrated that being part of the Big 4 and holding a foreign license are associated with higher transparency scores in these reports (Pheijffer, 2010; Deumes et al., 2012; La Rosa et al., 2019). Khlif and Achek (2017) highlighted a positive link between female representation in companies and greater social or environmental transparency. Van Buuren and Koch (2014) identified audit firm size as the primary determinant of the breadth and depth of transparency report disclosures across six EU countries. Smith (2021) noted that larger and sector-specialized audit firms tend to produce more readable audit reports. Demers et al. (2022) documented that institutional investor ownership is associated with longer and more positively-toned human resource disclosures. Wang et al. (2020) showed that textual emphasis varies in fundraising contexts and can attract investor attention differently. Li (2019) reported that firms increase repetitive disclosures when there is a new CEO or when introducing new information in notes. Given these findings, our research model posits that audit firm attributes play a crucial role in shaping the textual characteristics of transparency reports. Therefore, we propose our first hypothesis set as follows:

H1a: Audit firm-based characteristics have an influence on readability of TRs.

H1b: Audit firm-based characteristics have an influence on tone of TRs.

H1c: Audit firm-based characteristics have an influence on emphasis on basic auditing principles and best practices in TRs.

H1d: Audit firm-based characteristics have an influence on document intra-similarity of TRs.

H1e: AF-based characteristics have an influence on transparency score of TRs.

Deumes et al. (2012) examined the relationship between TR disclosure levels, audit firm characteristics, and audit quality, finding only a weak association between quality control scores and audit quality. In contrast to their cross-country approach, our study focuses on a single-country sample to clarify this relationship by minimizing international differences in regulations and contexts. Furthermore, we hypothesize that various textual attributes of TRs, such as tone, emphasis on basic principles (EoBP), document similarity, and readability, may also have a link to audit quality. Zeng et al. (2021) provided evidence by linking content attributes of key audit matters, such as readability and document length, to the likelihood of issuing modified audit opinions, a proxy for audit quality. TRs present an opportunity for audit firms to differentiate themselves (Girdar and Jeppesen, 2018), and Demers et al. (2022) demonstrated that profitable firms tend to make more original statements in their reports, while less successful firms rely on generic, less informative language. Therefore, we anticipate that the textual attributes of an AF's transparency report in a given year are indicative of the quality of audit performance during that period. In instances where TRs are generic and lack transparency, characterized by predominantly positive tones, we expect any significant effects on audit quality to be unlikely. Thus, we propose our next set of hypotheses as follows:

H2a: TR's readability has a role on audit quality.

H2b: TR's tone has a role on audit quality.

H2c: TR's emphasis on basic principles and best practices has a role on audit quality.

H2d: TR's document intra-similarity has a role on audit quality.

H2e: TR's transparency score has a role on audit quality

3. RESEARCH DESIGN and METHODOLOGY

3.1. Research Sample and Dataset

To construct our research sample, we initially obtained a list of AFs engaged in audit activities in the Turkish capital market between 2013 and 2020. This list was provided by the Capital Markets Board (CMB) of Türkiye. Subsequently, we visited the official websites of these AFs through the Public Oversight, Accounting, and Auditing Standards Authority (POAASA) website and downloaded all available TRs. Using the firm list of Public Disclosure Platform (PDP) website, we identified publicly traded firms as our client firm list. We then focused on the financial notifications of these firms on the PDP website by utilizing web scraping techniques, thereby creating a client firm-year based dataset which includes audit report parameters from these notifications. The detailed data extraction process will be outlined in section 3.2. Table 1 illustrates the methodology employed to create the sample for our study.

Table 1. Sample Creation Process

Panel A: Number of Unique AFs engaged in audits in Turkish capital market	
2013-2020	**106**
2013	92
2014	92
2015	98
2015	100
2017	99
2018	104
2019	104
2020	104
Number of audit partners working at these AFs.	1237
Number of unique audit partners in the research dataset	279
Number of TRs disclosed between 2013-2020*	612
Number of TR/AF-year based observations	612
Panel B: Steps to CF-Based Research Sample	
Number of firms on PDP websites	548
Number of firms in financial data by EquityRT	553
Number of matching CF-year based observations (2013-2020)	3525

*: In Türkiye, TR disclosure in a specific year is compulsory for only those firms audited a public interest entity in that year. Thus, all AFs does not have to prepare its TRs for every fiscal year.

This study utilizes several subdatasets. Firstly, the research sample comprises 612 TR-year observations spanning from 2013, the initial year of promulgation by the regulator, to 2020. Automated software methods were primarily applied for data processing in this dataset. Secondly, AF attribute dataset consists of AF-year observations, and these attributes were extracted from TRs. Sanctions information was obtained from Capital Markets Board (CMB) bulletins using automated web scraping methods. Thirdly, AP attribute dataset also includes AP-year observations, and these were likewise acquired from TRs. Fourthly, CFs' audit report characteristics were scraped from the Public Disclosure Platform (PDP) website. Lastly, CFs' financial data were sourced from the EquityRT financial data provider firm. Ultimately, we consolidated a panel dataset by aggregating all subdatasets, utilizing the data scraped from the Public Disclosure Platform (PDP) website since this dataset inherently encompasses AF-year, CF-year, and AP-year observations. The sources of these subdatasets are summarized in Table 2.

Table 2. Sources of Research Datasets

No	Dataset Class	Source
1	AF Attributes Dataset	TRs*
2	AP Attributes Dataset	TRs**
3	TR Attributes Dataset	TRs
4	CF Attributes Dataset	Public Disclosure Platform (PDP) and CF websites

5	CF Financial Data	EquityRT Financial Data Provider Firm
6	CF Audit Report Attributes Dataset	Public Disclosure Platform

*: Merely, capital market auditing activities license list and delist dates were provided by CMB of Türkiye. Also, sanctions were web-scraped through CMB Bulletins available on the Board's website.
**: Some AP attributes such as board membership were obtained from Central Securities Depository (CSD) of theTurkish capital markets.

This study utilizes several subdatasets. Firstly, the research sample comprises 612 TR-year observations spanning from 2013, the initial year of promulgation by the regulator, to 2020. Automated software methods were primarily applied for data processing in this dataset. Secondly, AF attribute dataset consists of AF-year observations and these attributes were extracted from TRs. Sanctions information was obtained from Capital Markets Board (CMB) bulletins using automated web scraping methods. Thirdly, AP attribute dataset also includes AP-year observations and these were likewise acquired from TRs. Fourthly, CFs' audit report characteristics were scraped from the Public Disclosure Platform (PDP) website. Lastly, CFs' financial data were sourced from the EquityRT financial data provider firm. Ultimately, we consolidated a panel dataset by aggregating all subdatasets, utilizing the data scraped from the Public Disclosure Platform (PDP) website since this dataset inherently encompasses AF-year, CF-year, and AP-year observations. The sources of these subdatasets are summarized in Table 2.

3.2. Measures

3.2.1. Audit Quality Measures as Dependent Variable

Various proxies have been used to measure audit quality. Following prior research, which considers GC and modified audit opinions as direct and widely used indicators of audit quality (DeFond and Zhang, 2014; Sun et al., 2020; Athavale et al., 2022; Bradbury and Kim, 2023), we used modified audit opinions as our proxy. Audit opinion data and audit report parameters were collected from the Public Disclosure Platform (PDP) via a Python-based Selenium web scraping bot. A dichotomous variable was then created, assigning 1 to qualified, adverse, and disclaimer opinions, and 0 to unqualified opinions.

Despite the criticism on prior research (please see Defond and Zhang, 2014), discretionary accruals has been the most prominent proxy for audit quality in auditing research. Amidst various models proposed in the literature (Healy, 1985; DeAngelo, 1986; Jones 1991; DeChow et al., 1995; Kothari et al., 2005), compatible with the recent research (e.g. Xiao et al., 2020, Garcia-Blandon et al., 2020), we also employed Standard Jones Model (Jones, 1991) and Modified Jones Model (DeChow et al., 1995). Since the financial data provided by EquityRT did not include the item "*changes in current portion of long-term debt*," we refrained from applying the Modified Jones Model with ROA (Kothari et al., 2005). Standard Jones Model (SJM) defines total accruals as specified in equation 1:

$$TA_{it} = \Delta CurrentAssets_{it} - \Delta Cash_{it} - \Delta CurrentLiabilities_{it} - \Delta DAE_{it}\,, \quad (1)$$

where TA_{it}= total accruals in year t for firm i, $\Delta CurrentAssets_{it}$= current assets in year t less current assets in year t-1, $\Delta Cash_{it}$= cash in year t less cash in year t-1, $\Delta CurrentLiabilities_{it}$ = current liabilities in year t less current liabilities in year t-1, and ΔDAE_{it}= change in depreciation and amortization expense between the years t and t-1.

In order to estimate the coefficients used in discretionary and non-discretionary accruals calculation, the model in equation 2 was employed:

$$\left(\frac{TA_{it}}{A_{it-1}}\right) = \alpha_0 + \alpha_1\left(\frac{1}{A_{it-1}}\right) + \beta_1\left(\frac{\Delta R_{it}}{A_{it-1}}\right) + \beta_2\left(\frac{PPE_{it}}{A_{it-1}}\right) + \in_{it}\,, \quad (2)$$

where TA_{it}= total accruals in year t for firm i, ΔR_{it}= change in revenues between the years t and t-1, PPE_{it}= gross fixed assets, plant, and equipment in year t, A_{it-1}= total assets in year t-1 and $\in_{it}$= error

term in year t.

Similar to the SJM, in order to estimate earnings management through Modified Jones Model (MJM), equation 3 was used:

$$TA_{it} = \Delta CurrentAssets_{it} - \Delta Cash_{it} - \Delta CurrentLiabilities_{it} + \Delta STD_{it} - \Delta DAE_{it}\ , \quad (3)$$

where, TA_{it}= total accruals in year t for firm i, $\Delta CurrentAssets_{it}$= current assets in year t less current assets in year t-1, $\Delta Cash_{it}$= cash in year t less cash in year t-1, $\Delta CurrentLiabilities_{it}$ = current liabilities in year t less current liabilities in year t-1, ΔSTD_{it}= debt included in current liabilities in year t less debt included in current liabilities in year t – 1, and ΔDAE_{it}= change in depreciation and amortization expense between the years t and t-1. The next step in MJM is presented in equation 4:

$$NA_{it} = \widehat{\alpha_0} + \widehat{\alpha_1}\left(\frac{1}{A_{t-1}}\right) + \widehat{\beta_1}\left(\frac{\Delta R_{it} - \Delta AR_{it}}{A_{t-1}}\right) + \widehat{\beta_2}\left(\frac{PPE_{it}}{A_{t-1}}\right) + \epsilon_{it}\ , \quad (4)$$

where NA_{it}= normal accruals in year t for firm i; ΔR_{it}= revenues in year t less revenues in year t-1, ΔAR_{it}= accounts receivables in year t less accounts receivables in year t-1, PPE_{it}= gross fixed assets, plant, and equipment in year t, A_{t-1}= total assets in year t-1, and εit = error term in year t for firm i. Lastly, in order to calculate discretionary accruals (DA_{it}), total accruals (TA_{it}) and normal accruals (NA_{it}) in hand, the equation 5 is needed:

$$DA_{it} = \frac{TA_{it}}{A_{it-1}} - NA_{it}\ , \quad (5)$$

Ultimately, absolute value of the residuals in these earnings management models represent the two other audit quality proxies in our research.

3.2.2. TR Attributes as Dependent and Independent Variables

3.2.2.1. Readability

To assess the readability of TRs, we employed the Fog Index, one of the most widely used readability measures in accounting and finance research due to its simplicity and replicability (Gangadharan and Padmakumari, 2024; Efretuei and Hussainey, 2023). The index is based on average sentence length and the proportion of complex words. Despite some criticism (Loughran and McDonald, 2014), prior studies emphasize its transparency and practical advantages (Demers et al., 2022). The Fog formula is presented in Equation 6.

$$0.4\ (\mathrm{ASL} + 100\frac{\mathrm{N_{wsy}}}{\mathrm{N_w}})\ , \quad (6)$$

Where, ASL = Average Sentence Length; N_wsy = Number of words with three or more syllables, and N_w = Total number of words.

In the meantime, we should notify that standard Fog scale is baseline readability scale in our study for follow-up studies to validate our study findings.

Given that the Gunning Fog Index was originally developed for English, we attempted to adapt it to Turkish, where word and sentence structures differ substantially. The Fog Index is based on average sentence length and the proportion of complex (polysyllabic) words. Prior studies argue that the English

threshold for complex words is not fully suitable even in financial contexts (Loughran and McDonald, 2014), and this limitation becomes more pronounced in Turkish, where average word and sentence lengths are generally higher (Atesman, 1997; Bezirci and Yilmaz, 2010). Supporting this view, Atesman (1997), who developed the Turkish adaptation of the Flesch Reading Ease Test, reported higher average syllable and sentence lengths in Turkish compared to English.

In an attempt to develop our Turkish Fog Scale, firstly, 114 texts which have corresponding Turkish and English versions were chosen from various fields at varying difficulty levels. These were 40 peer-reviewed articles on literature, philosophy, law, physics, 18 financial texts and 13 general culture texts, 17 short stories, 16 jokes, and 10 fairy tales. For English versions, Fog Index calculation tools available on web portal were used and these calculations were once again automated using Selenium by building a web scraping bot in Phyton. For those Turkish versions, scripts were created through Phyton, aligned with the grammar rules of counting syllables in Turkish. The syllable counting process has been iterated for all conceivable threshold values ranging from 1 to 8. In essence, this entails generating potential Fog index scores for 8 distinct complex word threshold values. Finally, we attempt to reach a coefficient that minimizes the difference between the English Fog scores and alternative Turkish Fog estimations. Ultimately, a unique equation was derived with a coefficient of 0.40 and a complex word threshold value of five syllables, as specified in Equation 7:

$$0.4\,(\mathrm{ASL} + 100\frac{\mathrm{N_{wsy}}}{\mathrm{N_w}})\,, \tag{7}$$

where ASL = Average Sentence Length; N_wsy = Number of words with five or more syllables; N_w = Total number of words. In contrast to the Fog Index, which indicates readability levels by grade, we established a class interval (Black, 2004, pp.22-24) utilizing the upper and lower limits of the experimental results from our sample of 114. This classification is based on five distinct classes (very hard, hard, medium, easy, very easy), similar to the approach employed by Atesman (1997). It ranges between 0 and 30 values. Hence, the intervals 24-30; 18-24; 12-18; 6-12; 0-6 correspond to the very hard, hard, medium, easy, and very easy, respectively. Ultimately, all members of the sample belonging to the easy/hard classes exhibited compatibility with their respective difficulty levels.

3.2.2.2. Tone

Following prior research (Demers et al., 2022; Teng and Han, 2022; Ehrmann and Talmi, 2020; Ertugrul et al., 2017), our investigation into the tone of TRs involved the utilization of a word list, specifically the LM Dictionary, tailored for financial texts by Loughran and McDonald (2011). This list encompasses 353 positive words, 2354 negative words, 296 uncertain (neutral) words, and 26 weak modal words. In the first place, like the studies in languages other than English (Teng and Han, 2022), these English words were translated into Turkish, while taking their synonyms in Turkish into consideration. In the translation process, we utilized Selenium framework to construct a web scraping bot on Google Translate, which offers a multilingual neural machine translation tool developed by Google. Additionally, we employed Pandas, a Phyton library for data analysis. The final word list, including synonymous words, comprises 599 positive, 4017 negative, 495 uncertain, and 48 weak modal words. In the sentiment analysis stage, the frequencies of these words were calculated. The initial LM dictionary is imbalanced across tone groups, and the addition of synonymous words during translation further contributed to this disproportionality. To address this, we implemented a normalization procedure on the dataset.

Given the highest number of words in the negative word category, we adjusted each word frequency by multiplying it with the ratio of negative word numbers to the total word numbers in that word's category (e.g., 4017/599 for positive words). Additionally, to account for document length, a second normalization procedure was conducted, dividing the scaled values in each TR document by the total number of words in that document. Finally, we calculated a relative variable based on the scaled frequencies of positive, negative, neutral, and weak modal words. Considering image management,

stakeholder theory, and signal theory, it is not anticipated that TRs would exhibit a negative tone. The expectation is that TRs, akin to other corporate disclosures, would likely convey a positive tone. Alternatively, due to audit firms' gravity and confidentiality, with the associated uncertainty about proprietary costs, TRs might express an uncertain or neutral tone. Following the approach of certain studies in the extant research (Ehrmann and Talmi, 2020, p.50), we derived a net positivism score over neutrality by subtracting the number of uncertain (neutral) words from the number of positive words in each document file. Finally, values greater than zero indicate a relatively positive tone, while values lower than zero suggest a relatively neutral tone.

3.2.2.3. Emphasis on Basic Audit Principles and Best Practices (EoBP)

Wang et al. (2020) demonstrated that the emphasis on investment decisions, present in various textual sources such as titles and detailed descriptions, significantly influences different aspects. Building on prior TR research and our own detailed TR reviews, we identified fifteen principles. These principles mainly encompass fundamental auditing principles emphasized by regulatory bodies, as well as best practices that have gained prominence in business environments. These are (1) audit quality, (2) ethics, (3) independence, (4) quality control or assurance system, (5) continuing education, (6) sustainability, (7) innovation (or improvement), (8) artificial intelligence (or machine learning, digital transformation), (9) client, (10) sanction, (11) regulatory/supervisory authority, (12) risk, (13) standard, (14) performance evaluation (or performance appraisal), and (15) Covid-19 (or pandemic, Coronavirus). To that end, first, all TRs text files and then fifteen keywords in JavaScript Object Notation (JSON) format were read in Phyton, and their frequencies were calculated. Factor analysis was conducted using STATA 15, employing varimax rotation without constraining the number of factors. Following Hair et al. (2010), factor loadings below 0.50 were disregarded, and a single principal factor explaining 79.4% of the variation was retained. Consequently, eight items remained loaded on this factor.

The Cronbach's Alpha for the eight-item scale was calculated as 82.03%, indicating strong evidence for the reliability of this new scale. This value exceeds the threshold of 0.70, as proposed by Nunnally and Bernstein (1994). Upon examination of the remaining six concepts that failed to load on one factor, it becomes evident that these are distinct and unrelated concepts. In contrast, the final eight items on the factor were composed of basic principles and best practices. We termed this new scale "Emphasis on Basic Principles and Best Practices-EoBP". Finally, as illustrated in Equation 8, we computed the scores for this new construct based on the factor analysis weights.

$$EoBP = 0.78\,AuditQuality + 0.49\,Ethic + 0.52\,Independence + 0.53\,Sustainability + 0.63\,Innovation + 0.62Client + 0.87Risk + 0.58\,Standards \quad (8)$$

3.2.2.4. Document Similarity

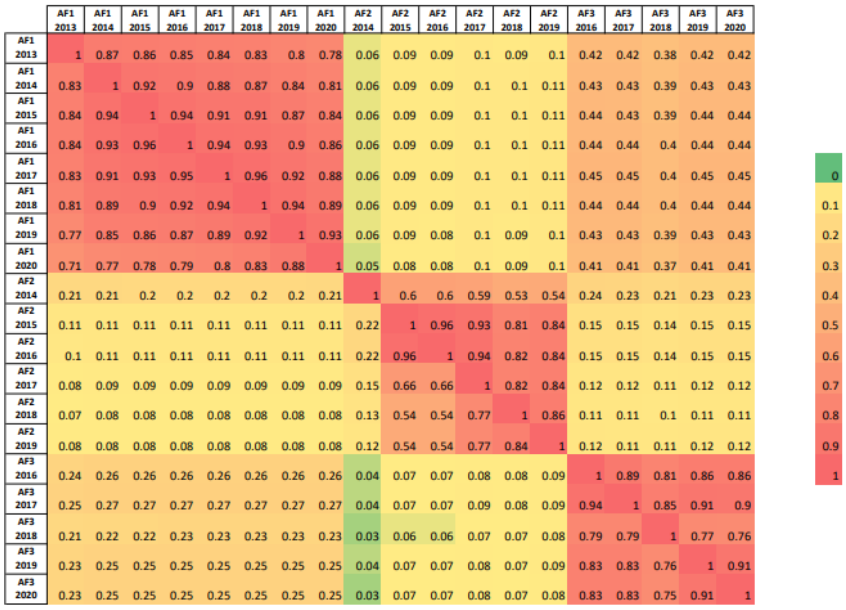
Examining the template or boilerplate effect in TRs requires measuring document similarity. Unlike studies using word-embedding approaches such as cosine similarity (Ehrmann and Talmi, 2020), we employed through WCopyFind[1], a plagiarism-detection-based tool that directly compares text files and measures the magnitude of similarity without transforming documents into vectors. This approach better aligns with our objective of identifying repetitive disclosures. While Li (2019) used another plagiarism detection software, MOSS, to examine repetitive disclosures in annual reports, WCopyFind was preferred due to its open-source structure.

All TRs in text format were loaded into WCopyFind, which uses n-gram similarity based on sequences of n consecutive words. After setting parameters (3-grams, no punctuation, and no case sensitivity), the program generated pairwise similarity scores for all TRs. The outputs were processed in Python using the Pandas library and visualized as a 612 × 612 heatmap matrix (provided as supplementary material).

[1] https://plagiarism.bloomfieldmedia.com/software/wcopyfind/

Figure 1 shows an anonymized example with three firms, where similarity scores range from 0.0 (completely different reports) to 1.0 (identical reports). It should be noted that similarity scores are not necessarily symmetric. For instance, AF2's 2015 vs. 2014 report similarity is 0.22, while the reverse comparison yields 0.60, suggesting substantial additions in 2015. In contrast, AF2's 2015 and 2016 reports show near-identical content, with a similarity score of 0.96 in both directions. These findings are discussed in detail in a later section.

Using this matrix, we defined a new distinct variable called *intra-similarity* as the average similarity score across a specific firm's TRs disclosed in different years, representing the document similarity in our study. The heat matrix also enables us to visually detect which AFs prepare boilerplate TRs most and share more common content with the other AFs, enriching our discussion on document similarity findings. No normalization procedure was needed since the WCopyFind program already gives a document's number of matched words by dividing by the number of words in that document, hence taking the document's length into consideration. Furthermore, we defined another distinct variable, referred to as *inter-similarity*. This variable represents the average similarity score between a specific firm's TR disclosed in a particular year, and TRs of all other firms across all years. However, this variable is not integrated into our research model. Nevertheless, it serves as a valuable metric, effectively illustrating the extent of shared content within a firm's transparency report in comparison to reports from other firms.



	AF1 2013	AF1 2014	AF1 2015	AF1 2016	AF1 2017	AF1 2018	AF1 2019	AF1 2020	AF2 2014	AF2 2015	AF2 2016	AF2 2017	AF2 2018	AF2 2019	AF3 2016	AF3 2017	AF3 2018	AF3 2019	AF3 2020
AF1 2013	1	0.87	0.86	0.85	0.84	0.83	0.8	0.78	0.06	0.09	0.09	0.1	0.09	0.1	0.42	0.42	0.38	0.42	0.42
AF1 2014	0.83	1	0.92	0.9	0.88	0.87	0.84	0.81	0.06	0.09	0.09	0.1	0.1	0.11	0.43	0.43	0.39	0.43	0.43
AF1 2015	0.84	0.94	1	0.94	0.91	0.91	0.87	0.84	0.06	0.09	0.09	0.1	0.1	0.11	0.44	0.43	0.39	0.44	0.44
AF1 2016	0.84	0.93	0.96	1	0.94	0.93	0.9	0.86	0.06	0.09	0.09	0.1	0.1	0.11	0.44	0.44	0.4	0.44	0.44
AF1 2017	0.83	0.91	0.93	0.95	1	0.96	0.92	0.88	0.06	0.09	0.09	0.1	0.1	0.11	0.45	0.45	0.4	0.45	0.45
AF1 2018	0.81	0.89	0.9	0.92	0.94	1	0.94	0.89	0.06	0.09	0.09	0.1	0.1	0.11	0.44	0.44	0.4	0.44	0.44
AF1 2019	0.77	0.85	0.86	0.87	0.89	0.92	1	0.93	0.06	0.09	0.08	0.1	0.09	0.1	0.43	0.43	0.39	0.43	0.43
AF1 2020	0.71	0.77	0.78	0.79	0.8	0.83	0.88	1	0.05	0.08	0.08	0.1	0.09	0.1	0.41	0.41	0.37	0.41	0.41
AF2 2014	0.21	0.21	0.2	0.2	0.2	0.2	0.2	0.21	1	0.6	0.6	0.59	0.53	0.54	0.24	0.23	0.21	0.23	0.23
AF2 2015	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.22	1	0.96	0.93	0.81	0.84	0.15	0.15	0.14	0.15	0.15
AF2 2016	0.1	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.22	0.96	1	0.94	0.82	0.84	0.15	0.15	0.14	0.15	0.15
AF2 2017	0.08	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.15	0.66	0.66	1	0.82	0.84	0.12	0.12	0.11	0.12	0.12
AF2 2018	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.13	0.54	0.54	0.77	1	0.86	0.11	0.11	0.1	0.11	0.11
AF2 2019	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.12	0.54	0.54	0.77	0.84	1	0.12	0.11	0.11	0.12	0.12
AF3 2016	0.24	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.04	0.07	0.07	0.08	0.08	0.09	1	0.89	0.81	0.86	0.86
AF3 2017	0.25	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.04	0.07	0.07	0.09	0.08	0.09	0.94	1	0.85	0.91	0.9
AF3 2018	0.21	0.22	0.22	0.23	0.23	0.23	0.23	0.23	0.03	0.06	0.06	0.07	0.07	0.08	0.79	0.79	1	0.77	0.76
AF3 2019	0.23	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.04	0.07	0.07	0.08	0.07	0.09	0.83	0.83	0.76	1	0.91
AF3 2020	0.23	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.03	0.07	0.07	0.08	0.07	0.08	0.83	0.83	0.75	0.91	1

Figure 1. Heat Matrix of TR Similarity Scores for three AFs spanning from 2013 to 2020 (0.0 is unrelated; 1.0 is perfect match)

3.2.2.5. Transparency Score

When calculating the transparency score of TRs spanning the years 2013-2020, our methodology involved initially creating an extensive question pool. This pool was generated through prior research, particularly the work of Van Buuren and Koch (2014), as well as insights from our dedicated readership of TRs and relevant regulations. Initially comprising 96 questions, we refined this pool to a final set of 88 items. The distribution of these questions is detailed in Table 3. To maintain brevity in this context, we have opted not to list all 88 questions here. However, the complete set is available upon request.

Table 3. Distribution of Question Pool

Section	# of Items
Cover Page and Design (CPD)	7
Legal Structure and Ownership (LSO)	12
Key Person and Partners (KPP)	9
Global Audit Network (GAN)	6
Associated AFs (AAF)	2
Organizational Structure (OS)	4
Quality Assurance Review (QAR)	5
Independence Compliance (IC)	4
Audited Public Interest Entities (APIE)	2
Continuing Education Policy (CEP)	6
Partners' Remuneration (PR)	8
Quality Control System (QCS)	13
Financial Information (FI)	5
Other Issues (OI)	1
General (G)	4
Total	**88**

Out of 88 questions, 35 were answered using a rule-based named entity recognition approach implemented via Python if–else blocks, selected for items with clear Yes/No answers. First, all TRs were converted to text using the Textract library, and Turkish characters were replaced with corresponding Unicode forms. The if–else scripts then coded responses as 1 (Yes) or 0 (No). Some questions were too abstract for automated extraction, and in some cases, answers were missing from text versions due to OCR/data loss in scanned reports. These 21 items were therefore manually coded by eight undergraduate research assistants to ensure accuracy. For the remaining 32 questions, a sample of 612 TRs (106 TRs, including at least one report per audit firm) was created. Using Notepad++, answers were first identified and labeled in this sample set, including their positions, and these labels were then used to develop Python scripts to search and extract the same information from the remaining 506 documents.

3.2.3. Audit Firm, Audit Partner Attributes, and Client Firm Controls

Previous studies show that audit firm characteristics influence disclosure quality in TRs, so we include these attributes in our model. Prior research finds that transparency is positively associated with audit firm size (Petersen and Zwirner, 2009), as well as Big 4 affiliation and foreign licensure (Pheijffer, 2010; Deumes et al., 2012; La Rosa et al., 2019). Firm size also emerges as a key determinant of TR disclosure breadth and depth across EU countries (Van Buuren and Koch, 2014). In addition, female representation is linked to higher levels of social and environmental transparency (Khlif and Achek, 2017), while larger and sector-specialized firms tend to produce more readable audit reports (Smith, 2021).

Based on prior auditing literature, we also include audit firm/audit partner (AF/AP) tenure (Stanley and DeZoort, 2007; Al-Thuneibat et al., 2011), AP board membership, gender (Breesch and Branson, 2009), and AP busyness (Karjalainen, 2011; Gul et al., 2013; Sundgren and Svanström, 2014; Cheng et al., 2021). Control variables include Altman's Z-score (Alcalde et al., 2022), CF age (Johl et al., 2021; Ertugrul et al., 2015), and CF sales ratio (Cheng et al., 2021), which are detailed below.

3.2.3.1. Audit Firm Attributes

License duration variable was employed as a proxy for AF age, where a higher license duration indicates the AF's older age. To calculate this variable, we determined the difference between the CMB list date and delist date or the difference between the list date and the current date if the AF still holds a license for conducting audit activities in capital markets. Subsequently, the natural logarithm of the number of

days was incorporated into the research model. *Sanctions* variable was obtained using computer-based techniques. Initially, all Capital Markets Board (CMB) Bulletins available on the CMB website[2] were systematically downloaded using Selenium, a powerful open-source automation tool that leverages the Chrome webdriver. This was achieved by constructing a web scraping bot and utilizing the Request[3] library in Python. The CMB imposes two types of sanctions on AFs: (1) Administrative fines, measured in monetary terms, and (2) the cancellation of independent audit authorization, denoted as deauthorization. Administrative fines are consistently presented in tabulated form within a specific section of the bulletin, facilitating extraction using the Camelot[4] module. Conversely, deauthorization is articulated within sentences containing specific phrases, and these keywords were identified using the PDFMiner module. Both sets of sanction data were then consolidated into an Excel file. Given the disparate scales and the limited occurrence of administrative fines and audit deauthorization, we combined them and subsequently transformed the data into a dummy variable.

The remaining AF attributes aside from license duration and sanctions, were gathered manually from TRs due to the unavailability of longitudinal data in Türkiye. *Big-4* is a dummy variable, taking the value 1 for Deloitte, EandY, KPMG, PwC, and 0 for other firms. The *partners* variable signifies the number of audit partners in an AF. *Ownership structure* is represented by the highest capital share held by audit partners in an AF, indicating concentrated ownership with higher values and relatively dispersed ownership with lower values, following prior research (Goergen and Renneboog, 2001, p. 266). *Foreign license* is a dummy variable that takes the value 1 if an AF possesses any foreign licensure and 0 otherwise. *Female partner ratio* serves as a proxy for female representation, determined by identifying the gender of audit partners through their names extracted from TRs. For names with common usage for both genders or those that are less common, we conducted searches on the partners' LinkedIn accounts. In rare instances where profiles were absent, direct inquiries were made to AFs. The ratio was then computed based on AFs. *Non-audit services (NAS)* dummy variable was derived from financial revenue information in TRs. Firstly, both audit and non-audit service incomes in cash were recorded in Excel. It takes the value of 1 if the AF distributes non-audit services, and 0 otherwise.

In the final step, we computed additional attributes for AFs on a yearly basis on concatenated data. The *AF tenure* represents the number of successive years an AF has audited any CF. This variable was transformed by taking its natural logarithm. Regarding *AF market share*, it is defined as the number of CFs audited by the AF in a specific year divided by the total number of CFs audited by all AFs in that year.

3.2.3.2. Audit Partner Attributes

As mentioned above, *AP gender* was obtained manually from TRs since the data were not available through regulatory and supervisory bodies. *AP board membership (BM)* data were sourced from the Central Securities Depository (CSD) and Trade Repository of Türkiye. Additionally, *AP tenure* and *AP busyness* were computed on a yearly basis on concatenated data. In this context, *AP tenure* represents the number of successive years during which any audit partner audited any client firm, and the variable was transformed by taking its natural logarithm. *AP busyness* denotes the total number of audits conducted by a specific audit partner in a given year.

3.2.3.3. Control Variables

Altman Z-score, first developed by Altman (1968), is a multivariate bankruptcy prediction model used to assess firms' financial distress risk. Despite criticism, it remains widely applied in finance, banking, and credit risk studies and is often considered one of the most effective bankruptcy prediction tools (Marsenne et al., 2024; Alves and Meneses, 2024; Alcalde et al., 2022; Al-Manaseer and Oshaibat,

[2] https://spk.gov.tr/spk-bultenleri/gecmis-yillara-ait-bultenler.
[3] https://realpython.com/python-requests/.
[4] https://pypi.org/project/camelot-py/.

2018; Calandro, 2007). It is also commonly used by practitioners such as bankers, investors, fund managers, and credit rating agencies. Although several revisions exist, the original model is still widely used, while recent studies suggest that Altman's final logistic specification provides a more appropriate estimation of failure probability (Alcalde et al., 2022). Accordingly, we employ the final logistic model in Equations 9 and 10.

$$Z = \frac{1}{1 + e^{-\mathrm{L}}} \tag{9}$$

$$\begin{aligned} L = {} & -13.302 - 0.459X_1 - 1.160\,X_2 - 1.682\,X_3 - 0.013\,X_4 - 0.034\,D_1 \\ & - 0.150D_2 - 0.631D_3 + 1.837\,S_1 - 0.061S_2 + 0.186\,A_1 \\ & - 0.099A_2 - 0.628\,I_1 + 0.365\,I_2 - 0.157\,I_3 - 0.176\,I_4 + 0.095\,I_5 \\ & - 0.472\,I_6 - 0.915\,I_7 - 0.014\,C_1\,, \end{aligned} \tag{10}$$

where X_1= Working Capital/Total Assets; X_2= Accumulated Reserves/Total Assets; X_3= EBIT/Total Assets; $X_{4=}$ Market Value/Total Assets. $D_1 - D_3$ represents the years of 2008, 2009, 2010 respectively hence we did not include these dummies into our computation due to the study sample between the years 2013-2020. S_1 and S_2 refer to natural logarithms of total assets and the square of total assets, respectively. A_1 and A_2 notate age dummies as less than 6 years and over 12 years, respectively. We take I_1 = 1 for restaurants and hotels, I_2 = 1 for building, I_3 = 1 for wholesale and retailing, I_4 = 1 for agriculture, I_5 = 1 for manufacturing, I_6 = 1 for energy and water production, and I_6 = 1 for information technology. Among the model parameters, we excluded year and country risk dummies from our logistic model estimation. This decision stems from the fact that our study adopts a single-country research design and encompasses the sample period of TRs from 2013 to 2020.

Moreover, *age*, *sales-to-total assets ratio*, and *loss* variables were incorporated into the model as controls for CFs. In the case of *CF age*, the firm's foundation year was identified by searching the firm's website or the footnotes of financial statements within embedded financial notifications through the Public Disclosure Platform (PDP) website. The other two financial ratios were generated using a financial dataset provided by EquityRT, a financial data provider. Table 4 provides a comprehensive overview of all variables included in both the main model estimation stages and the robustness check. Lastly, in order to observe the impact of COVID in transparency reports, we investigated whether there is a mention of COVID in them. If mentioned, we created a dummy variable assigning a value of 1; otherwise, a value of 0.

Table 4. Variable Definitions

Variable	Definition
MAO	1 if the audit opinion is adverse, qualified or disclaimer of the opinion, otherwise 0.
SJM	Absolute value of the residuals of Standard Jones Model.
MJM	Absolute value of the residuals of Modified Jones Model.
Sanctions	1 if CMB imposed administrative fines or cancellation of independent audit authoritization against AFs in relevant year; 0 otherwise.
Fog	0.40 x (average sentence length + 100 x complex words*/total words).
Turkish Fog	0.40 x (average sentence length + 100 x complex words**/total words).
EoBP	Factor score of the EoBP construct with eight items.
Tone	(Number of normalized positive words/Total number of words)-(Number of normalized uncertain words/Total number of words).
Document Similarity	The average of intra-similarity scores among AF's TRs disclosed in different years as a proxy for boilerplate effect.
Inter-Similarity	The average similarity score between a specific firm's TR disclosed in a particular year and TRs of all other firms across all years (not included into research model).
Transparency Score	The natural logarithm of transparency score obtained through the answers of 88 questions.

TR Timeliness	The number of days between legal term regarding fiscal year to disclose TRs and AF's TR disclosure date.
License Duration	The natural logarithm of license duration in terms of number of days as a proxy for AF age.
Partners	The number of partners in AFs.
Big-4	1 if AF belongs to Deloitte, EandY, KPMG, PwC global accounting network; 0 otherwise.
Ownership Structure	Highest percentage of shares in an AF in relevant year.
Female Partners	The ratio of the number of female partners to total number of partners in relevant year.
NAS	1 if AF distributes non-audit services; 0 otherwise.
Foreign License	1 if AF has a foreign licensure; 0 otherwise.
Sanctions	1 if AF is fined by CMB of Türkiye in a specific year; 0 otherwise.
Number/Letter	Number of numbers/ Number of letters in TRs.
AF Market Share	Number of CFs for an AF in a specific year/ Number of CFs audited by all AFs in a specific year.
AF Tenure	Natural logarithm of the successive years an AF audited a particular CF.
AP Gender	1 if AP is male; 0 otherwise.
AP BM	1 if AP is a member of board; 0 otherwise.
AP Busyness	Natural logarithm of number of audits conducted by a specific AP in a given year.
AP Tenure	Natural logarithm of the successive years which an AP audited a particular CF.
CF Age	Natural logarithm of the number of years since the establishment of a particular client firm.
CF Sales Ratio	Net Sales/ Total Assets ratio of CF.
CF Z Score	Altman's logistic Z score value, referring to the likelihood of a CF to fail.
CF Loss	1 if CF experienced a loss for a specific year; 0 otherwise.
TR Covid Effect	1 if Covid, Covid 19, pandemic, Coronavirüs were mentioned in TRs; 0 otherwise.

*: the number of words with equal to greater than 3 syllables.
**: the number of words equal to greater than 5 syllables.

3.3. Analysis

3.3.1. Preliminary Analysis Before Model Estimation

Firstly, for all datasets, whether collected manually or obtained through software methods, we conducted a thorough data cleaning process to eliminate the noise in the data. Subsequently, we ranked all subsets to identify outliers and address potential measurement errors. Consequently, these errors were corrected. Secondly, concerning the Turkish Fog scale, we encountered a single instance of a negative readability score. While mathematically possible, it lacks conceptual interpretability. Hence, we removed the readability score of this report from consideration, resulting in a final dataset with 611 available observations in TR-year based data.

Excluding the outliers mentioned above, we chose not to implement any data modification or outlier treatment processes such as winsorized mean or trimmed mean, for three main reasons. Firstly, the few outliers identified represent the actual values of the variables. For instance, the TR timeliness variable includes two outliers, -243 and 214, which correspond to significantly delayed and unusually early disclosure dates by audit firms. Likewise, the discretionary accrual variables display some extreme values, which represent the absolute residuals derived from estimating both the Standard Jones and Modified Jones Models. These values reflect inherent variability in practice. Essentially, these models are already crafted to assess abnormal accruals and fluctuations in earnings management practices across firms. Hence, as stated by Foorthuis (2021, p. 297), apart from being noise, these outlier values might represent the signals researchers are looking for. Additionally, techniques such as winsorization and trimming could be robust approaches, however, they may lack efficiency (Abdullahi, 2024, pp. 2-8). Further, we retained these outliers resulting from the inherent variability of the data, a practice supported by prior research (Anscombe and Guttman, 1960; Osborne, 2002; Osborne and Overbay, 2004). Secondly, several variables underwent a transformation process. Altman's Z score, discretionary accruals, and certain other variables were obtained through model estimation procedures. The natural logarithm was applied to some variables (e.g., license duration, AF tenure, AP tenure). Word occurrences for estimating reports' tone were processed twice for normalization and scaling. EoBP was

already a distinct construct obtained through principal component analysis, computed from the loadings of eight factors. Some metric variables were transformed into dichotomous equivalents (i.e., client firm's loss). These data transformation, rescoring, and coding processes align with methods recommended by prior research (see Osborne and Overbay, 2004; p. 3) to preserve actual outliers in the model. Table 5 provides descriptive statistics for all variables, whether included in the research models or not.

Table 5. Descriptive Statistics

Variable	N	Mean	St. Dev.	Min	Max
MAO	3524	.106	0.308	0	1
Standard Fog	3183	31.061	2.232	24	40
Turkish Fog	3183	12.710	1.752	8	19
EoBP	3183	118.523	71.253	7.669	285.078
Tone	3183	6.467	13.575	-27.446	43.628
Doc.(Intra) Similarity	3179	0.659	0.147	0.126	0.955
Doc. (Inter) Similarity	3183	0.059	0.032	0.015	0.150
Transparency Score	3183	36.140	3.490	20	45
License Duration	3183	8.990	0.415	6.836	9.458
Partner	3174	16.267	8.252	1	27
Female Partners	3174	0.230	0.156	0	0.5
Ownership structure	3080	0.345	0.230	0.071	1
NAS	3181	0.975	0.158	0	1
Sanctions	3183	0.149	0.356	0	1
Foreign License	3183	0.918	0.274	0	1
Number/Letter	3183	0.021	0.009	0.004	0.072
AF Tenure	3512	1.380	0.503	0	2.079
AP Gender	3524	0.807	0.395	0	1
AP Board Membership	3142	0.588	0.492	0	1
AP Busyness	3524	1.469	0.766	0	2.996
AP Tenure	3524	0.930	0.543	0	1.792
CF Age	3516	3.306	0.727	0	4.771
CF Sales/TA	2741	1.178	5.111	-0.027	174.632
CF Z Score	2654	0.814	0.292	0	1
CF Loss	3515	0.224	0.417	0	1
TR Covid	3183	0.060	0.238	0	1
TR Timeliness	3075	12.578	39.929	-243	214
AF Market Share	3524	0.091	0.071	0.002	0.213

3.3.2. Pooled Regression Model Estimation with Robust Standard Errors (VCE)

To examine AF-based determinants of all TR textual attributes and address potential heterogeneity issues, we utilized a pooled regression model with robust standard errors, some of which are detailed in our previous studies ((Ozcan et al., 2023; Aydemir et al, 2022; Aydemir et al., 2021). For the sake of brevity, we present the results in a model comparison on Table 6. The research model at this stage is outlined below:

$$\mathrm{TR_{i,t}} = \beta_0 + \sum_{k=1}^{j} \beta_{1:j}\, \mathrm{AF_{i,t}} + \sum_{j+1}^{m} \beta_{j+1:m}\ Control_{i,t} + \varepsilon_{i,t}\,, \tag{11}$$

where $TR_{i,t}$ = TR textual attributes of audit firm i in year t, $AF_{i,t}$= Firm-based attributes of audit firm i in the year t, $Control_{i,t}$ = Control variables of in year t, and $\varepsilon_{i,t}$= error term of the model.

Table 6. Model Comparison

TR Attribute	Readability		EoBP	Tone	Intra Similarity	Transparency
Variable	Std. Fog	Turkish Fog				
License Dur.	**-0.351***	**-0.616*****	**6.678***	-0.790	**-0.051*****	0.016
	(0.182)	(0.170)	(3.499)	(1.056)	(0.010)	(0.012)
Partner	**-0.052****	**-0.074*****	-0.118	0.007	0.000	-0.001
	(0.023)	(0.022)	(0.360)	(0.119)	(0.002)	(0.002)
Big 4	**2.639*****	**2.017*****	**178.041*****	**-20.686*****	**-0.178****	0.076
	(0.869)	(0.754)	(19.505)	(5.721)	(0.069)	(0.057)
Ownership	**1.622*****	**0.658***	-0.450	**-8.869*****	**0.063***	-0.015
	(0.440)	(0.378)	(7.073)	(2.463)	(0.032)	(0.038)
Foreign Lic.	0.317	**0.500****	**11.171*****	-1.040	0.023	-0.013
	(0.251)	(0.219)	(3.583)	(1.370)	(0.016)	(0.018)
Sanctions	0.309	0.373	**18.389***	-0.931	-0.010	0.028
	(0.353)	(0.288)	(9.651)	(1.979)	(0.025)	(0.020)
Female ratio	**-1.595****	-1.254	**-28.013****	-1.780	-0.050	-0.002
	(0.766)	(0.762)	(12.015)	(4.344)	(0.043)	(0.042)
NAS	**-0.840*****	**-0.562***	**11.321*****	0.335	-0.031	0.021
	(0.294)	(0.300)	(3.715)	(1.516)	(0.021)	(0.033)
2014	0.480	0.248	2.574	-0.719	0.040	0.013
	(0.535)	(0.431)	(5.362)	(3.147)	(0.042)	(0.038)
2015	-0.170	-0.196	10.901	-0.608	**0.115*****	0.014
	(0.499)	(0.403)	(7.079)	(2.850)	(0.030)	(0.037)
2016	0.552	0.238	**18.650*****	-0.203	**0.111*****	0.049
	(0.498)	(0.394)	(6.369)	(2.774)	(0.032)	(0.034)
2017	0.207	-0.177	**18.093*****	0.020	**0.128*****	**0.058***
	(0.479)	(0.386)	(6.095)	(2.838)	(0.032)	(0.035)
2018	-0.004	-0.191	**24.28*****	-0.390	**0.104*****	0.053
	(0.487)	(0.384)	(6.471)	(2.748)	(0.031)	(0.034)
2019	0.409	0.062	**23.604*****	-0.577	**0.108*****	**0.063***
	(0.484)	(0.392)	(7.003)	(2.729)	(0.031)	(0.035)
2020	0.437	0.169	**23.266*****	-0.263	**0.093*****	**0.058***
	(0.491)	(0.390)	(6.330)	(2.722)	(0.032)	(0.035)
TR Timeliness	**-0.003****	**-0.004*****	**-0.065****	**0.032****	0.000	**0.000***
	(0.001)	(0.001)	(0.028)	(0.015)	(0.000)	(0.000)
Number/Letter	**-99.300*****	**-50.930*****	**-958.390*****	35.310	**-1.140***	**1.350****
	(9.980)	(8.060)	(139.910)	(61.360)	(0.600)	(0.620)
Market Share	**-22.530*****	**-9.010***	**-463.730*****	22.390	0.350	0.393
	(6.160)	(5.070)	(127.340)	(43.950)	(0.495)	(0.410)
Constant	37.950***	20.510***	-4.500	26.120**	1.100***	3.290***
	(1.675)	(1.544)	(30.518)	(10.149)	(0.092)	(0.104)
RSquare	*0.416*	*0.267*	*0.595*	*0.183*	*0.287*	*0.13*
F	*16.585*	*10.089*	*18.814*	*6.410*	*6.410*	*4.866*
N	*355*	*355*	*355*	*355*	*353*	*355*

*** p<.01, ** p<.05, * p<.1, standard errors in parenthesis.

The model comparison table above reveals significant findings. Firstly, similar results are observed for both the standard and Turkish Fog scales, with the exception of female ratio and foreign license. Across both scales, it can be inferred that the readability of TRs improves when AFs have a longer license duration, a greater number of partners, provide non-audit services, possess a larger market share, exhibit higher TR timeliness, and present a higher number-to-letter ratio in TRs. However, TRs become more challenging to read when AFs have Big 4 involvement and higher ownership concentration. Thus, H1a is supported for both readability scales.

Regarding tone, the results indicate that both Big 4 firms and firms with uneven ownership distribution exhibit a relatively neutral tone compared to others. However, TR tone tends to be more positive when

TR timeliness is higher, thus providing support for H1b. EoBP is notably higher for audit firms with a longer license duration, Big 4 involvement, foreign license, non-audit services, and sanctions imposed by the Capital Markets Board of Türkiye. Conversely, EoBP tends to decrease when the audit firm has a higher female ratio, greater TR timeliness, a higher number-to-letter ratio in TRs, and a larger market share. Consequently, H1c is supported.

The document similarity (boilerplate effect) is significantly lower for Big 4 firms and audit firms with a longer license duration and a greater number-to-letter ratio in TRs. Conversely, it is higher for firms with higher ownership concentration, aligning with the pertinent hypothesis, H1d. Lastly, the findings indicate that none of the explanatory variables, except for the number-to-letter ratio, is significant in explaining the variation in the transparency score. The transparency score is substantially higher for AFs with a greater number-to-letter ratio in TRs. Since the number-to-letter ratio, representing our control variable in these models, is not an attribute specific to audit firms, it can be concluded that H1e is not supported.

The audit firm market share was identified to exert a negative influence on TR readability for both scales. This result holds significance as it implies that TRs become more readable when the AF market share is higher, aligning with stakeholder theory. Similarly, the emphasis on basic audit principles tends to decrease when the audit firm has a higher market share, while it increases when the audit firm is fined by the Capital Markets Board of Türkiye in that year, indicating a focus on image management. On the other hand, year effects are not found to be significant for both readability scales, although there is a noticeable trend indicating an increase in EoBP from 2016 to 2020. In terms of intra-similarity score, the year effect is significantly positive across all years, indicating a substantial rise in the boilerplate effect over time. Additionally, the year effect is significant and positive only for 2017, 2019, and 2020, contributing to an enhanced transparency level.

3.3.3. Instrumental Variables Probit Model Estimation

The results from the model above, combined with insights from our previous studies (Ozcan et al., 2023; Aydemir et al, 2022; Aydemir et al., 2021) provided evidence for the role of AF-based attributes on TR textual attributes. To address the endogeneity issue, we employed the Instrumental Variables Probit Model with the two-step option. We used TR textual attributes to link to the likelihood of audit firms issuing modified audit opinions, which serves as a robust proxy for audit quality. Considering their significant roles in influencing TR attributes, license duration, number of partners, non-audit services (NAS), foreign license, and number-to-letter ratio were selected as instrumental variables. However, ownership structure, the Big 4 involvement, female partner ratio, and market share variables—although found to be significant in our prior models—were not included as instrumental variables due to their higher correlation scores with the variables in this model, thereby drawing attention to the issue of multicollinearity. The research model in this stage is as follows:

$$\mathrm{AQ_{i,t}} = \beta_0 + \sum_{k=1}^{j} \beta_{1:j}\,\mathrm{AF_{i,t}} + \sum_{j+1}^{m} \beta_{j+1:k}\,\mathrm{AP_{i,t}} + \sum_{m+1}^{n} \beta_{k+1:m}\mathrm{TR\_Fitted}_{\,i,t} + \sum_{n+1}^{s} \beta_{m+1:s}\,\mathrm{Controls_{i,t}} + \varepsilon_{i,t}\,, \tag{12}$$

where $AQ_{i,t}$= Audit quality of audit firm i in the year t, $AF_{i,t}$= Firm-level attributes of audit firm i in the year t, $AP_{i,t}$= Partner-level attributes of audit firm i in the year t, $\mathrm{TR\_Fitted}_{i,t}$= Predicted values of TR textual attributes of audit firm i in the year t, $Control_{i,t}$ = Control variables of in year t, and $\varepsilon_{i,t}$= error term of the model.

We utilized both the standard Fog scale and our newly developed Turkish Fog scale in the Instrumental Variables Probit Model. Consequently, Tables 7 and 8 display the results of the first and second stages

of this model using the standard Fog readability scale, while Tables 9 and 10 present the equivalent outcomes for the new Turkish Fog scale. For consistency in reporting, we adhered to the approach used in previous studies employing the IV Probit model, such as Colak et al. (2021). In the initial step of the analysis, we generated predicted or fitted values of TR textual attributes using the selected instrumental variables, considering these variables as endogenous. Table 7 shows that the F-values of all textual attributes indicate statistical significance in the first stage models for all endogenous content attributes. The models exhibit higher explanatory power, ranging from nearly 27% to 53%. Furthermore, our selected instruments are statistically significant, and the directions of their effects are consistent with the findings from our previous studies mentioned above.

Table 7. First Stage Regression Results of IV Probit Model (Standard Fog)

Variables	***Readability (Fog)***		***EoBP***		***Tone***	
	β_j	**St. Err.**	β_j	**St. Err.**	β_j	**St. Err.**
License Duration	**-1.111***	0.100	**-12.764**	2.743	-1.072	0.677
Partner	**-0.115**	0.005	**2.550**	0.150	**-0.706**	0.037
NAS	**-0.574**	0.222	**36.705**	6.069	-1.665	1.498
Foreign License	0.040	0.134	**24.057**	3.669	**-4.875**	0.906
Number/Letter	-70.852	3.878	**-1405.208**	105.599	**183.755**	26.079
AF Tenure	0.018	0.088	-1.318	2.406	-0.857	0.594
AP Gender	**0.214**	0.102	**-12.063**	2.778	0.463	0.686
AP BM	**-0.827**	0.081	**-39.918**	2.211	**-1.424**	0.546
AP Busyness	**-0.578**	0.049	1.744	1.350	**-0.659**	0.333
AP Tenure	**-0.451**	0.080	1.546	2.189	**-1.694**	0.540
CF Age	**-0.194**	0.057	0.823	1.561	-0.009	0.385
CF Sales/TA	-0.002	0.008	0.002	0.230	-0.054	0.056
CF Z Score	0.184	0.134	**-31.333**	3.662	1.266	0.904
CF Loss	-0.019	0.078	-3.432	2.148	0.383	0.530
TR Covid	**0.951**	0.203	**51.770**	5.536	1.157	1.367
Year	*Yes*		*Yes*		*Yes*	
N	*2346*		*2346*		*2346*	
F(22,2323)	*82.360*		*117.550*		*49.750*	
Rsquare	*0.4329*		*0.5223*		*0.3138*	
	Intra Similarity		***Transparency Score***			
	β_j	**St. Err.**	β_j	**St. Err.**		
License Duration	**-0.059**	0.006	**-0.015**	0.005		
Partner	**-0.007**	0.000	**0.004**	0.000		
NAS	**-0.031**	0.014	-0.020	0.011		
Foreign License	0.001	0.008	0.006	0.006		
Number/Letter	**-0.482**	0.252	**1.115**	0.200		
AF Tenure	-0.000	0.005	0.007	0.004		
AP Gender	**0.021**	0.006	**-0.022**	0.005		
AP BM	-0.001	0.005	-0.005	0.004		
AP Busyness	**-0.008**	0.003	**0.020**	0.002		
AP Tenure	0.005	0.005	-0.004	0.004		
CF Age	-0.006	0.003	0.002	0.002		
CF Sales/TA	0.000	0.000	-0.000	0.000		
CF Z Score	**0.021**	0.008	**-0.028**	0.006		
CF Loss	-0.000	0.005	0.005	0.004		
TR Covid	**0.025**	0.0132	**0.128**	0.010		
Year	*Yes*		*Yes*			
N	*2346*		*2346*			
F(22,2323)	*81.530*		*39.340*			
Rsquare	*0.4304*		*0.2645*			

*: Significant effects at p< 0.05 are in bold fond type.

Table 8 presents the results of the second stage probit model, including the marginal coefficients of the explanatory and control variables. The Wald test chi-square statistic (110.71) and its significance level

(0.000) provide substantial evidence supporting the overall goodness-of-fit of the model. Additionally, below the table, the Wald test's endogeneity chi-square statistic tests the null hypothesis that there is no endogeneity. The significant value (20.49) led to rejecting the null hypothesis, indicating evidence of endogeneity. Had we failed to reject it, it would have been suggested that the standard probit model might be more appropriate than the model that considers TR attributes as endogenous variables with their selected instruments. Therefore, this result reinforces the appropriateness of the IV probit model[5].

We also conducted a weak-instrument-robustness test for the IV estimation of the probit model, and the Wald chi-square statistic (29.66) was found to be significant. This statistic provides evidence that the coefficients of endogenous variables are different from zero, indicating that the instruments are not weak. Examining the marginal effects, the results reveal a statistically significant association between TR readability and the likelihood of issuing modified audit opinions. Specifically, audit firms that produce more readable TRs tend to issue more modified audit opinions, providing evidence for higher audit quality. Moreover, AF tenure, AP busyness, and CF sales are found to be negatively related to audit quality. The likelihood of issuing modified audit opinions is higher in years when the CF incurs a loss. The year effect is also significant across all years, indicating that audit firms tend to issue fewer modified audit opinions over time. Consequently, H2a is supported, while H2b, H2c, H2d, and H2e, which propose notable links between tone, emphasis on basic principles, document similarity, transparency score, and audit quality, are rejected.

Table 8. Second Stage Results of IV Probit Model (Standard Fog)

Variables	***Marginal Effects***	***Sta. Err.***	***z***	***P>z***	***95% Confidence Interval***	
Fog	**-0.423**	0.177	-2.390	0.017	-0.770	-0.076
EoBP	0.007	0.007	0.970	0.332	-0.007	0.023
Tone	0.018	0.056	0.330	0.741	-0.092	0.129
Transparency S.	-4.433	3.728	-1.190	0.234	-11.740	2.873
Document S.	8.576	4.832	1.770	0.076	-0.895	18.048
AF Tenure	**-0.367**	0.113	-3.240	0.001	-0.589	-0.145
AP Gender	-0.133	0.200	-0.670	0.505	-0.525	0.259
AP BM	-0.365	0.415	-0.880	0.378	-1.179	0.448
AP Busyness	**-0.162**	0.081	-2.000	0.045	-0.322	-0.003
AP Tenure	-0.179	0.208	-0.860	0.388	-0.588	0.229
CF Age	0.047	0.075	0.620	0.534	-0.101	0.196
CF Sales/TA	**-0.252**	0.066	-3.790	0.000	-0.382	-0.122
CF Z Score	0.119	0.322	0.370	0.712	-0.512	0.750
CF Loss	**0.407**	0.098	4.130	0.000	0.213	0.601
TR Covid	0.251	0.785	0.320	0.749	-1.288	1.791
2014	**-1.111**	0.581	-1.910	0.056	-2.251	0.028
2015	**-1.688**	0.814	-2.070	0.038	-3.285	-0.093
2016	**-2.074**	0.790	-2.620	0.009	-3.624	-0.525
2017	**-2.079**	0.862	-2.410	0.016	-3.770	-0.389
2018	**-1.713**	0.887	-1.930	0.053	-3.452	0.025
2019	**-1.527**	0.759	-2.010	0.044	-3.016	-0.038
2020	**-1.037**	0.507	-2.040	0.041	-2.032	-0.042
N	2346					
Wald ChiSquare	110.710					
Sign.	0.000					

Wald's Endogenity Test (Chi square=20.49, p=0.001)

In Table 9 below, the findings regarding the new Turkish Fog scale indicate that license duration, number of partners, and foreign license have a significant influence on our endogenous Turkish Fog scale. Almost all instrumental variables exhibit a significant effect on the TR tone attribute. The effects of license duration, number of partners, and NAS on document similarity are significant, while only

[5] https://www.stata.com/manuals/rivprobit.pdf.

license duration and the number of partners have a significant effect on the endogenous transparency score variable. Therefore, all first-stage regression results with F values greater than 10 suggest that the selected instruments are not weak and are relevant for our endogenous TR textual attributes, aligning with the criteria proposed by Staiger and Stock (1997).

Table 9. First Stage Regression Results of IV Probit Model (Turkish Fog)

Variables	***Readability (Turkish Fog)***		***EoBP***		***Tone***	
	β_j	St. Err.	β_j	St. Err.	β_j	St. Err.
License Duration	**-1.169**	0.086	**-12.765**	2.743	-1.072	0.677
Partner	**-0.074**	0.005	**2.551**	0.150	**-.0706**	0.037
NAS	-0.242	0.189	**36.705**	6.070	-1.665	1.498
Foreign License	**0.263**	0.114	**24.058**	3.669	**-4.875**	0.906
Number/Letter	**-36.072**	3.294	**-1405.208**	105.600	**183.755**	26.079
AF Tenure	0.016	0.075	-1.318	2.407	-0.857	0.594
AP Gender	0.097	0.087	**-12.063**	2.779	0.463	0.686
AP BM	**-0.714**	0.069	**-39.919**	2.212	**-1.424**	0.546
AP Busyness	**-0.254**	0.042	1.745	1.350	**-.659**	0.333
AP Tenure	**-0.258**	0.068	1.547	2.190	**-1.694**	0.540
CF Age	**-0.131**	0.049	0.823	1.562	-0.009	0.385
CF Sales/TA	-0.005	0.007	0.003	0.230	-0.054	0.056
CF Z Score	-0.140	0.114	**-31.334**	3.662	1.266	0.904
CF Loss	-0.037	0.067	-3.433	2.149	0.383	0.530
TR Covid	**1.130**	0.173	**51.770**	5.536	1.157	1.367
Year	*Yes*		*Yes*		*Yes*	
N	2346		2346		2346	
F(22,2323)	53.820		117.550		49.750	
Rsquare	0.3314		*0.5223*		0.3138	
	Intra-Similarity		***Transparency Score***			
	β_j	St. Err.	β_j	St. Err.		
License Duration	**-0.059**	0.007	**-0.015**	0.005		
Partner	**-0.007**	0.000	**0.004**	0.000		
NAS	**-0.032**	0.014	-0.020	0.011		
Foreign License	0.002	0.009	0.006	0.007		
Number/Letter	-0.482	0.252	**1.116**	0.200		
AF Tenure	-0.001	0.006	0.007	0.005		
AP Gender	**0.022**	0.007	**-0.022**	**0.005**		
AP BM	-0.002	0.005	-0.006	0.004		
AP Busyness	**-0.008**	0.003	**0.021**	0.003		
AP Tenure	0.005	0.005	-0.005	0.004		
CF Age	-0.006	0.004	0.002	0.003		
CF Sales/TA	0.000	0.001	0.000	0.000		
CF Z Score	**0.021**	0.009	**-0.028**	0.007		
CF Loss	0.000	0.005	0.006	0.004		
ŞR Covid	**0.026**	0.013	**0.128**	0.010		
Year	*Yes*		*Yes*			
N	2346		2346			
F(22,2323)	81.530		39.340			
Rsquare	*0.4304*		*0.2645*			

*: shows significance at p<.05, in bold font type.

In Table 10 below, the second-stage results of the IV probit model are presented. The Wald test chi-square statistic (53.87) and its significance level (0.0002) provide strong evidence supporting the overall goodness-of-fit of the model. Below the table, Wald's endogeneity chi-square statistic tests support the presence of endogeneity. Therefore, this result also suggests that the IV probit model is suitable and appropriate.

Table 10. Second Stage Results of IV Probit Model (Turkish Fog)

Variables	***Marginal Coeff.***	***Sta. Error***	***Z***	***P>z***	***95% Confidence Interval***	
Turkish Fog	-1.303	0.929	-1.400	0.161	-3.125	0.518
EoBP	0.020	0.016	1.260	0.208	-0.011	0.052
Tone	0.028	0.091	0.320	0.751	-0.150	0.208
Transparency S.	1.336	4.039	0.330	0.741	-6.580	9.253
Document S.	22.150	17.259	1.280	0.199	-11.677	55.979
AF Tenure	**-0.360**	0.188	-1.920	**0.055**	-0.730	0.008
AP Gender	-0.117	0.327	-0.360	0.719	-0.760	0.525
AP BoM	-0.372	0.683	-0.550	0.586	-1.713	0.968
AP Busyness	-0.276	0.182	-1.510	0.131	-0.634	0.082
AP Tenure	-0.383	0.467	-0.820	0.412	-1.301	0.533
CF Age	0.019	0.123	0.160	0.873	-0.222	0.261
CF Sales/TA	**-0.258**	0.069	-3.740	**0.000**	-0.394	-0.123
CF Z Score	0.129	0.514	0.250	0.801	-0.879	1.138
Loss	**0.382**	0.158	2.410	**0.016**	0.072	0.693
TR Covid	-0.420	1.121	-0.380	0.707	-2.619	1.778
2014	-3.630	2.638	-1.380	0.169	-8.802	1.542
2015	-5.355	3.886	-1.380	0.168	-12.973	2.262
2016	-5.779	3.827	-1.510	0.131	-13.282	1.722
2017	-6.322	4.320	-1.460	0.143	-14.790	2.146
2018	-5.372	3.989	-1.350	0.178	-13.193	2.447
2019	-4.926	3.559	-1.380	0.166	-11.903	2.051
2020	-3.316	2.356	-1.410	0.159	-7.935	1.303
N	2346					
Wald ChiSquare	53.870					
Sign.	0.0002					

Wald's Endogenity Test (Chi square= 23.51; p=0.0003)

We also performed a weak-instrument-robustness test for the IV estimation of the probit model, and the Wald chi-square statistic (12.84) is significant (p=0.0249). This statistic provides evidence that the coefficients of endogenous variables are different from zero and indicates that the instruments are not weak. Regarding the marginal effects, AFs that prepare more readable TRs tend to issue more modified audit opinions, although this effect is not significant for the new Turkish Fog scale. AF tenure and CF sales are negatively related to modified audit opinions. These opinions are also likely to be issued by AFs in years when the CF has incurred a loss. Year effects were found to be insignificant.

In general, the analyses support the appropriateness of the suggested model, which addresses the endogeneity issue by treating TR content attributes as endogenous regressors. The findings from the instrumental variable probit model indicate that only TR readability significantly influences the likelihood of an audit firm issuing a modified audit opinion, serving as a proxy for audit quality. Notably, AF tenure, CF sales ratio, and CF loss are identified as statistically significant variables, aligning with expectations from existing auditing research. Specifically, AF tenure exhibits a negative impact on the probability of a modified audit opinion. As the number of successive years that an audit firm audits a CF's financials increases, the likelihood of issuing a modified audit opinion tends to decrease. This suggests a potential risk to audit independence with prolonged audit relationships. Moreover, an increase in the CF sales-to-total-assets ratio is associated with a decreased probability of receiving a modified audit opinion, indicating a positive relationship between financial stability and audit quality. Conversely, if the CF incurs a loss in a specific year, the likelihood of receiving a modified audit opinion increases, reflecting the impact of financial performance on audit outcomes. Surprisingly, the analysis indicates that the effect of Covid on TRs does not yield statistically significant results, implying that the pandemic did not have a significant influence on the content of transparency reports.

Overall, our hypothesis that textual attributes in transparency reports influence audit quality was only substantiated for TR readability. However, the comprehensive results indicated that TR textual attributes

serve as endogenous regressors, and the selected instruments were found to be significant and not weak. This suggests that these features (i.e., instruments), many of which were individually proven to be significant in prior research on audit quality, also exert indirect effects on audit quality through their influence on TR textual attributes.

3.3.3. Robustness Check

3.3.3.1. Other Audit Quality Measures

To assess robustness, we examined whether TR textual attributes are associated with alternative audit quality proxies, namely discretionary accruals measured by the Standard Jones Model (SJM) and Modified Jones Model (MJM), using an IV 2SLS approach due to the continuous nature of the dependent variables. While first-stage regressions showed strong explanatory power, post-estimation tests indicated no evidence of endogeneity for either specification. However, the results reveal limited explanatory power of the model for discretionary accruals. In the SJM model, only audit partner busyness and client firm loss are significant, whereas no variables are significant in the MJM specification.

These findings are not surprising, as TR textual attributes are constructed at the audit firm level, and modified audit opinions also reflect audit firm-level audit quality. Thus, the primary model using modified audit opinions provides a more coherent framework for capturing the relationship between TR textual attributes and audit quality. In contrast, discretionary accruals are driven mainly by client firm behavior, weakening the conceptual link with audit firm-level textual characteristics. This aligns with prior literature distinguishing audit quality proxies, where DeFond and Zhang note that modified audit opinions are more direct and less error-prone, whereas accrual-based measures involve greater noise and measurement error.

While the accrual-based robustness tests are weak, our research model stands as a robust and comprehensive framework, encompassing audit firm and audit partner attributes along with client firm controls, as emphasized frequently in prior auditing research. We notably incorporate AF/AP tenure (Stanley and DeZoort, 2007; Al-Thuneibat et al., 2011), AP board membership and gender (Breesch and Branson, 2009), AP busyness (Karjalainen, 2011; Gul et al., 2013; Sundgren and Svanström, 2014; Cheng et al., 2021), and the study revealed comparable findings with the extant research. Furthermore, we incorporate control variables such as Altman's Z score (Alcalde et al., 2022), CF age (Johl et al., 2021; Ertugrul et al., 2015), as well as CF sales ratio (Cheng et al., 2021). It is noteworthy that our study aligns with existing research by revealing comparable findings for several variables, establishing connections with the extant literature.

4. DISCUSSION

This research aims to assess the readability, emphasis on basic audit principles and best practices, tone, document similarity, and transparency score of transparency reports published by 106 distinct Turkish audit firms. These firms have conducted independent audits on public interest entities and possess licenses to perform audits in the capital market. The study covers the transparency reports disclosed during the period from 2013 to 2020. The research employs a computer-based approach, a methodology predominantly applied to corporate disclosures or financial texts but not extensively employed in the analysis of transparency reports. Secondly, it intends to identify probable audit firm-related determinants of all these textual attributes. Lastly, it attempts to link them to modified audit opinions as a proxy for audit quality. According to DeFond and Zhang (2014, p. 285), modified audit opinions, compared to other audit quality measures, are recognized as more direct, more egregious, and actual indicators with lower measurement error, on which scholars also have more consensus. Therefore, we included it as the main dependent variable in our baseline research model.

Consequently, the average readability score of transparency reports based on the standard Fog scale (32.166) suggests that these documents are difficult to read, whereas our new Turkish Fog scale indicates an average value of 13.265, pointing to a medium level of readability. The Fog scale score of 32.166 is notably higher than averages reported for financial texts in the literature. For instance, Loughran and McDonald (2014) reported an average score of 18.94 for annual reports, while Kim et al. (2018) found an average score of 20 for 10-K filings. In the expanded auditor reporting setting, Smith (2021) reported scores between 24 and 29. Although the standard Fog scale serves as the baseline readability scale in our study for follow-up research to validate our findings, we tailored the Fog scale specifically to the Turkish language. Through an experiment aimed at optimizing our objective function, we developed a unique equation with the same coefficient (i.e., 0.40) but a threshold of five syllables as the complexity threshold for words. This threshold value is also consistent with Turkish studies, which note longer words in terms of syllables due to Turkish being an agglutinative language (Atesman, 1997).

EoBP is a scale derived from factor analysis of fifteen basic audit principles and best practices. One factor comprising eight principles has explained most of the common variance with a high scale reliability value (approximately 0.80), exceeding the acceptable threshold of 0.70 (Nunnally and Bernstein, 1994). Emphasis on basic audit principles is found to vary across audit firms, highlighting its distinctive attribute. Therefore, in the model identifying audit firm-based determinants of textual attributes, it exhibits the highest explanatory power with an R-square value approaching 60%. The results concerning transparency reports' tone indicate a predominantly positive tone, consistent with prior research on other corporate communication channels such as central bank monetary policy decisions (Ehrmann and Talmi, 2020), annual reports (Teng and Han, 2022), and human capital disclosures (Demers et al., 2022). Audit firms are known for their commitment to seriousness and confidentiality. Therefore, we were interested in whether their communications might lean towards a neutral tone. However, transparency reporting serves as a critical channel through which audit firms communicate with stakeholders. In alignment with stakeholder theory, it appears that they tend to favor a positive tone in their reports.

The document intra-similarity scores of audit firms themselves suggest that transparency reports are more boilerplate than documents conveying unique, distinctive information as designed and envisioned by standard setters. The minimum score is nearly 13%, while the maximum score is almost 96%, with an average of nearly 66%. Interestingly, one audit firm, neither from the Big 4 nor the second-tier, achieved the minimum intra-similarity score (12.6%). However, audit firms seem to have consistently disclosed uniform transparency reports between 2013 and 2020, contributing to dysfunctional reporting practices that hinder the reduction of information asymmetry. In elucidating this significantly higher intra-similarity score, we posit that audit firms may have encountered a trade-off. On one hand, they grapple with the challenge of furnishing comprehensive and accurate information regarding their internal governance and organizational structure, aiming to mitigate agency costs, alleviate information asymmetry, and enhance transparency. On the other hand, there is a reluctance to disclose proprietary information that could potentially impact their competitive advantage in the audit market. Ultimately, it appears that they opt to disclose boilerplate documents, striking a balance between providing necessary information and safeguarding strategic details.

Furthermore, inter-similarity scores (ranging from nearly 2% to 15%) among audit firms' transparency reports indicate that reports from different firms share common content to some extent, including exactly matching expressions. This finding can be partially explained by the fact that almost all audit firms mention Quality Control Standard 1 released by POAASA. The heat matrix, provided as supplementary material, indicates that audit firms share common content not with the Big 4, but rather with second-tier or other foreign licensed firms. Particularly, local audit firms that disclosed only one transparency report between 2013 and 2020 appear to have higher shared common content with another local audit firm, similar to itself, rather than with other audit firms. The Big 4 have smaller similarity scores for both intra- and inter-similarity compared to other audit firms, suggesting that they prepare relatively more original and unparalleled transparency reports. Moreover, one of the Big 4 has the minimum inter-similarity score (1.5%). In general, rather than preparing distinctive reports, some audit firms seem to prefer using general content from other transparency reports.

Unfortunately, the findings regarding the transparency scores of these reports reveal that the maximum score on the 88-item transparency scale is 45 points, indicating that audit firms are not as transparent as intended. Furthermore, the minimum score was 20, and the transparency scores of transparency reports do not appear to vary significantly across audit firms. This finding is remarkable but not surprising given the overwhelming similarity findings. Unfortunately, these reports are not only boilerplate documents but also fall short of enhancing transparency. The transparency scores of audit firms with and without foreign licenses are observed to be nearly identical.

The study's findings on audit firm-based determinants of textual attributes reveal significant results. Both the standard and Turkish Fog scales exhibit consistent patterns, except for female representation and foreign licenses. The presence of female auditors has a negative impact on standard scale readability, resulting in more readable reports, whereas foreign licenses positively affect the Turkish Fog scale, leading to less readable reporting. Overall, transparency report readability improves with longer license duration, more partners, non-audit services, larger market share, higher timeliness, and a higher number-to-letter ratio. Conversely, readability decreases for Big 4 firms and those with higher ownership concentration, consistent with prior research that emphasizes their detailed, complex reports due to extensive global accounting network references (Ehlinger, 2010; Fu et al., 2015). Year effects do not show significance for readability scales.

Emphasis on basic principles is notably higher for firms with longer license duration, Big 4 involvement, foreign licenses, non-audit services, and sanctions imposed by the Capital Markets Board of Türkiye. Conversely, it tends to decrease if the audit firm has a higher female ratio, timely reporting, a higher number-to-letter ratio in reports, and greater market share. There are two interesting findings here. First, audit firms that receive a sanction in a specific year emphasize audit principles and best practices more in that year's report. Second, those with a larger market share accentuate these concepts less. These findings strongly support the evidence of image management efforts in these reports. There is also a notable trend of increased emphasis on basic principles during the years 2016-2020. Regarding tone, the Big 4 firms and firms with uneven ownership distribution exhibit a relatively neutral tone compared to others. The tone tends to be more positive when audit firms publish their transparency reports earlier.

The boilerplate effect (intra-similarity) is notably lower for the Big 4 and firms with longer license duration, as well as those with a greater number-to-letter ratio in their reports, while it is higher for firms with higher ownership concentration. The Big 4 appear to prepare their annual transparency reports with relatively more original content. The boilerplate effect is consistently positive across all years, indicating a substantial increase in intra-document similarity over time. Unfortunately, this finding represents a significant obstacle in mitigating information asymmetry. Lastly, the findings reveal that none of the explanatory variables, except for the number-to-letter ratio, are significant in explaining the variation in transparency scores. To explain these nonsignificant findings, we argue that the transparency scores of reports (ranging between 20 and 45) do not actually vary significantly across audit firms; therefore, the heterogeneity of this textual attribute is not sufficient to differentiate one firm from another. Interestingly, the transparency score is substantially higher for audit firms with a higher number-to-letter ratio in their reports. In other words, firms that provide more numerical content or tables in their reports are also observed as more transparent. The year effect is also significant and positive for the years 2017, 2019, and 2020, indicating an increasing trend in transparency levels over time.

The findings investigating the link between textual attributes and modified audit opinions reveal some significant but mixed results for the models using two readability scales. We included the standard Fog scale in our main model to enhance the generalizability of our study findings and facilitate repeatability by future research. Therefore, suggesting textual attributes as endogenous variables, we estimated an instrumental variable probit model. Model statistics show that instruments are not weak and are linked to the endogenous regressors. The second-stage probit analysis results produced a statistically significant link between readability and modified audit opinions. Specifically, audit firms that prepare more readable reports tend to issue more modified audit opinions, providing evidence of higher audit quality. Furthermore, audit firm tenure, audit partner busyness, and client firm sales are negatively related to

modified audit opinions. Audit firms tend to issue more modified audit opinions in years when the client firm incurs a loss. The year effect is significant for all years, indicating that audit firms tend to issue fewer modified audit opinions over time. The results regarding audit firm tenure and audit partner busyness are consistent with previous research that reported a negative effect of audit firm tenure (Al-Thuneibat et al., 2011) and partner busyness (Sundgren and Svanström, 2014) on audit quality. Actually, existing research has produced controversial evidence. While some studies have shown that audit firm tenure negatively affects audit quality (Al-Thuneibat et al., 2011), others have indicated either a weak relationship (Knechel and Vanstraelen, 2007) or no significant relationship (Jenkins and Velury, 2012) between the two. The strong argument underlying this link is that the likelihood of an audit firm adhering to independence standards decreases if it has conducted successive audits for a client firm, a point supported by our study as well.

Moreover, our study aligns with previous research, showing that modified audit opinions tend to decrease when a client firm has a higher sales-to-total-assets ratio (Gul et al., 2013) and increase when the firm incurs a loss during the fiscal year (Kumar and Lim, 2015). While other textual attributes such as tone, emphasis on basic audit principles, document similarity, and transparency level do not show significant effects, our study provides compelling evidence for the significant association of a specific textual attribute -readability- with audit quality. Deumes et al. (2012) previously examined the impact of another textual attribute, disclosure level, on audit quality in a cross-country research design but did not find a significant association. In contrast, our research contributes by demonstrating a significant link between a textual attribute and audit quality within a single-country sample. This is achieved by mitigating cross-country differences and employing an instrumental variables probit model that treats transparency report textual attributes as endogenous regressors.

When evaluating the findings, some of which are quite interesting, audit firms appear to view transparency reports as a channel for communication with stakeholders and engage in specific image management efforts. For instance, in years when they receive a sanction from the Capital Markets Board of Türkiye, there is an increased emphasis on basic audit principles or best practices. However, audit firms with high market share tend to place less emphasis on these aspects in their reports.

In explaining the nonsignificant effects of other textual attributes on audit quality, such as emphasis on basic principles, tone, document similarity, and transparency level, we consider the hypotheses of image management and incomplete revelation. Image management efforts may have hindered the direct link between tone, emphasis on principles, and audit quality. Consequently, the expected relationship might have been obscured by noise stemming from image management strategies. Similarly, due to the proprietary cost associated with providing transparent and accurate information, the transparency score is notably low in the reports, while the boilerplate effect is significantly higher. It seems that the reports often disclose idealistic, ex-ante information rather than current and factual data, as suggested by Van Buuren and Koch (2014). Therefore, the hypothesized connections between document similarity or transparency score and audit quality may not have materialized. Importantly, the findings indicate a lack of differentiation in transparency scores across audit firms. Regardless of whether they are part of the Big 4, hold a foreign license, or have a senior or junior age profile, all audit firms seem to exhibit low transparency scores. Consequently, such homogeneous textual characteristics may struggle to have a meaningful impact on audit quality.

The findings from our research model using the new Turkish Fog scale indicate that all textual attributes are endogenous, and the selected instruments are not weak. However, the second-stage results of the IV probit model, using the fitted values of textual attributes, show that none of these attributes significantly influence modified audit opinions. Nevertheless, the effects of audit firm tenure and two factors related to client firms -sales ratio and financial loss in the current year- are consistent with both our main model and prior research. Specifically, audit firms conducting successive audits for the same client and those with higher client sales ratios tend to issue fewer modified audit opinions. Conversely, client firms experiencing financial losses are more likely to receive modified audit opinions. Audit partner busyness and the year effect were not found to be significant in this model. Additionally, in both models using

the standard and Turkish Fog scales, the effect of Covid-19 was not statistically significant.

We further assessed the robustness of our model using alternative audit quality proxies commonly employed in the audit literature. Instead of modified audit opinions, we applied the Standard Jones Model and Modified Jones Model as audit quality proxies from earnings management perspectives. The results from the Instrumental Variable Two-Stage Least Squares analysis indicate that textual attributes were not found to be endogenous in these models. Therefore, we did not find evidence supporting a significant role of textual attributes in influencing either discretionary accruals model. These results align with arguments from existing research and our research design considerations. Defond and Zhang (2014) highlight in their comprehensive review that discretionary accruals models are less direct proxies for audit quality compared to modified audit opinions. This is because audit firms have relatively limited influence over the discretionary accruals of client firms. They also note the susceptibility of these measures to higher levels of measurement error and bias, along with a lack of consensus in their application. In our study, we treat textual attributes as features at the audit firm level. Modified audit opinions, issued directly by audit firms, establish a more direct and logical connection with audit quality. Conversely, discretionary accruals models are influenced more by client firm behaviors, over which audit firms have less control. Therefore, our findings do not reveal a significant association between textual attributes and discretionary accruals models.

The additional findings from the Standard Jones and Modified Jones Models are consistent with our base model and prior research. For example, the Standard Jones Model shows that client firms tend to engage in higher earnings management when audit partners are more occupied, suggesting a negative impact of audit partner workload on audit quality. This indicates that auditors may face challenges in detecting abnormal earnings management practices in client firms during busy periods. Similarly, both models demonstrate that earnings management practices are more pronounced (indicating lower audit quality) when client firms report a loss for the year. These findings underscore the robustness of our primary model and reinforce our confidence in its validity.

This research offers valuable insights for regulatory and supervisory institutions, audit firms, and auditors. The findings indicate that transparency reports, despite being rich data sources, are often complex and fail to effectively reduce information asymmetry. Due to their insufficient transparency, these reports appear inadequate in fulfilling their mission to enhance transparency effectively. Therefore, regulatory bodies should consider these findings when shaping future policies. It is recommended that these reports use more accessible language and provide concrete, specific information to promote increased transparency, rather than offering general, repetitive content.

Here, we would like to offer our suggestions to the Public Oversight, Accounting and Auditing Standards Authority (POAASA) in Türkiye. Firstly, global authorities advocate for the use of plain, clear, and simple language in corporate disclosures. For example, the Securities and Exchange Commission (SEC) in the United States published *'A Plain English Handbook'* in 1998, aimed at improving corporate reports (see Bonsall et al., 2017, p.330). Therefore, POAASA in Türkiye could adopt similar guidelines to specify the language and style to be used in these reports. Secondly, transparency reports in Türkiye are uploaded to POAASA's information system, but many are in scanned file formats or use non-standard character sets. Unfortunately, this complicates converting these reports into text format and applying OCR techniques. This finding can be considered as an implicit indication that these reports have not been analyzed in-depth using computer-based methods by POAASA in Türkiye, or that the possibility of such an approach has not been taken into consideration. Therefore, the finding highlights the urgent need for more effective and sustainable reporting and monitoring mechanisms. Additionally, audit firms produce these reports in diverse types and formats, despite their similar content structure. Given this situation, we recommend POAASA consider implementing a standardized template to eliminate the presentation of transparency reports in varying formats. Thirdly, we urge POAASA in Türkiye to rigorously review transparency reports. These reports, envisioned by standard setters, aim to enhance transparency, reduce information asymmetry, and provide reliable information on audit processes and internal governance. However, our study findings reveal a prevalent

boilerplate effect and lower transparency levels in these reports. Consequently, they currently appear inadequate in achieving these critical objectives.

In countries where investor protection is weaker compared to developed nations, the analysis of transparency reports through computer-based and/or artificial intelligence methods, as in this study, plays a significant role in the adoption and institutionalization of sustainable reporting practices. It provides crucial insights to regulatory and supervisory bodies like POAASA in Türkiye and transforms into a powerful tool for strong monitoring and policymaking activities. Additionally, the findings of this study highlight the need for POAASA in Türkiye to develop policies to counteract the image management signals of audit firms, as the study provides evidence that audit firms frame the language in their transparency reports. For instance, the findings indicate that emphasis on basic auditing principles is higher in firms that have received sanctions, while it is relatively lower for audit firms with a higher market share. These reports should not turn into image management instruments, contrary to their intended purpose.

5. CONCLUSION

This study applies computer-based text analysis methods, including natural language processing, pattern recognition, named entity recognition, document similarity, and web scraping, to examine transparency reports of 106 Turkish audit firms. The results show that these reports are largely boilerplate, with moderate-to-high readability and a generally positive tone. We also document variation across firms and over time in the emphasis on basic audit principles.

Audit firm and audit partner characteristics significantly affect textual attributes such as readability, tone, emphasis on audit principles, document similarity, and transparency scores, confirming the endogenous nature of these features. Among these, only readability shows a meaningful association with audit quality measured by modified audit opinions, suggesting that more readable reports are linked to higher audit quality.

Overall, transparency reports appear to function more as communication tools than as effective mechanisms for reducing information asymmetry. Despite regulatory intent, they provide limited distinctive content, indicating a gap between expected and actual reporting outcomes. This highlights the need for improved reporting standards and monitoring practices. Future research may extend the analysis to annual or integrated reports, employ more advanced semantic techniques, and include report preparer characteristics to better understand reporting behavior.

ACKNOWLEDGEMENTS

This study has been supported by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the 1001 Funding Program, Grant No. 121K325, as part of the project titled "Do Audit Firms Disclose Enough: A Computer-Based Content Analysis of Turkish Audit Firms' Transparency Reports."

We also express our sincere thanks to Capital Markets Board of Türkiye for providing us with authorisation dates for audit firms to engage in audits in capital markets. We are also thankful to Central Securities Depository and Trade Repository of Türkiye for sharing the audit partners' board membership data with us.

REFERENCES

2006/43/EC of The European Parliament and of The Council. (2006). On statutory audits of annual accounts and consolidated accounts, amending Council Directives 78/ 660/EEC and 83/349/EEC and repealing Council Directive 84/253/EEC. Official Journal of the European Union, 88-107.

Abdullahi, K. B. (2024). Statistical mirroring: A robust method for statistical dispersion estimation. *MethodsX*, 12, 102682.

Alcalde, R., C. A Armino and S. Garcia (2022). Analysis of the economic sustainability of the supply chain sector by applying the Altman Z-score predictor. *Sustainability*, 14, 1-13.

Al-Manaseer, S. and S. D. Al-Oshaibat (2018). Validity of Altman Z-score model to predict financial failure: Evidence from Jordan. *International Journal of Economics and Finance*, 10(8), 181-189.

Al-Thuneibat, A., R. Tawfiq Ibrahim Al Issa and R. A.Ata Baker (2011). Do audit tenure and firm size contribute to audit quality? Empirical evidence from Jordan. *Managerial Auditing Journal*, 26(4), 317-334.

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23 (4), 589-609.

Alves, C.F. and L. L. Meneses (2024). ESG scores and debt costs: Exploring indebtedness, agency costs, and financial system impact. *International Review of Financial Analysis*, 94, 103240.

Anscombe, F.J. and I. Guttman (1960). Rejection of outliers. *Technometrics*, 2 (2), 123-147.

Arena, C., S. Bozzolan and G. Michelon (2015). Environmental reporting: Transparency to stakeholders or stakeholder manipulation? An analysis of disclosure tone and the role of the board of directors. *Corporate Social Responsibility and Environmental Management, John Wiley and Sons*, 22(6), 346-361.

Article 36-Transparency Reports (Independent Audit By-Law) 2012 (Public Oversight Accounting and Auditing Standards Authority-POAASA).

Atesman, E. (1997). Türkçe'de okunabilirliğin ölçülmesi. *Dil Dergisi*, 58, Ağustos 1997, 71-74.

Athavale, M., Z. Guo, Y. Meng and T. Zhang (2022). Diversity of signing auditors and audit quality: Evidence from capital market in China. *International Review of Economics and Finance*, 78, 554-571.

Aydemir S.D., Y. Kütük and M. Aksu (2021). Computer-Based Content Analysis of Audit Firms' Transparency Reports and Their Audit Firm Related Determinants. MODAVICA 2021 18th International Conference on Accounting, New Dynamics of Accounting-Transformation in Theory and Practice, November 25-27, 2021, Virtual.

Aydemir, S., Y. S. Akgül and M. Aksu (2022). The Role of Audit Firm Characteristics in Communicative Value of Transparency Reports: A Textual Analysis, 41th Eurasian Business and Economics Society Conference, October, 12-14, 2022, Berlin-Germany.

Bezirci, B. and A. E. Yilmaz (2010). Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazılım kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü, *DEÜ Mühendislik Fakültesi Fen Bilimleri Dergisi*, 12 (3), 49-62.

Black, K. (2004). Business statistics for contemporary decision making. 4th Edition, Leyh Publishing, LLC, USA.

Blanco, B., P. Coram, S. Dhole and P. Kent (2021). How do auditors respond to low annual report readability? *Journal of Accounting and Public Policy*, 40(3).

Bodnaruk, A., T. Loughran and B. McDonald (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4), 623-646.

Bonsall, S.B., A. J. Leone, B. P. Miller and K. Rennekamp (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63, 329-357.

Bozcuk, A. (2018). Türkiye'deki bağımsız denetim kuruluşlarında üst düzeydeki cinsiyet çeşitliliği ve gelir etkisi, *Uluslararası Sosyal Araştırmalar Dergisi*, 11(59), 882-890.

Bradbury, M.E. and O. Kim, (2023). The impact of the IFRS adoption reform on audit market concentration, auditor choice and audit quality, *Journal of Applied Accounting Research*, Vol. ahead-of-print No. ahead-of-print.

Breesch, D. and J. Branson (2009). The effects of auditor gender on audit quality. *Journal of Accounting Research and Audit Practices*. 8 (3/4), 78-107.

Calandro, J. (2007). Considering the utility of Altman's Z-score as a strategic assessment and performance management tool. *Strategy and Leadership*, 35(5), 37-43.

Cheng, Y., C. M. Haynes and M. D. Yu (2021). The effect of engagement partner workload on audit quality. *Managerial Auditing Journal*, 36(8), 1068-1091.

Cho, C., G. Michelon and D. M. Patten (2012). Impression management in sustainability reports: An empirical investigation of the use of graphs. *Accounting and the Public Interest*, 12(1), 16-37.

Cho, C.H., R. W. Roberts and D. M. Patten (2010). The language of US corporate environmental disclosure. *Accounting, Organizations and Society*, 35(4), 431-443.

Colak, G., D. Gounopoulos, P. Loukopoulos and G. Loukopoulos (2021). Tournament incentives and IPO failure risk. *Journal of Banking and Finance*, 130, 106193.

Cornell Law School. (2023, September 18). Legal Information Institute. Retrieved September, 18, 2023, from https://www.law.cornell.edu/wex/sarbanes-oxley_act).

Cular, M. (2009). Quality analysis of the annual reports in Croatia. Graduate Work, University of Split, Faculty of Economics, Split.

Cular. M. (2014). Transparency of AFs in Croatia. World Academy of Science, Engineering and Technology, *International Journal of Economics and Management Engineering*, 8 (3), 694-698.

DeAngelo, L. E. (1986). Accounting numbers as market valuation substitutes: A study of management buyouts of public stockholders. *The Accounting Review*, 61, 400-420.

DeChow, P. M., R. G. Sloan and A. P. Sweeney (1995). Detecting earnings management. *The Accounting Review*, 70(2), 193-225.

DeFond, M.L. and J. Zhang (2014). A review archival auditing research. *Journal of Accounting and Economics*, 58 (2-3), 275-326.

Demers, E., V. X. Wang and K. Wu (2022). Corporate human capital disclosures: Early evidence from the SEC's disclosure mandate. SSRN

Deumes, R., C. Schelleman, H. V. Bauwhede and A. Vanstraelen (2012). Audit firm governance: Do Transparency Reports reveal audit quality? *Auditing: A Journal of Practice and Theory*, 31(4), 193-214.

Dyer, T., M. Lang and L. Stice-Lawrence (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64, 221-245.

Efretuei, E. and K. Hussainey (2023). The fog index in accounting research: contributions and challenges. *Journal of Applied Accounting Research*, 24 (2), 318-343.

Ehlinger, A. (2010). Determinants of the extent of disclosure in Transparency Reports according to Article 40 of the 8th company law directive: First evidence from Austria, Germany and the Netherlands. The Maastricht Accounting, Auditing and Information Management Research Center, Aachen, 46-72.

Ehrmann, M. and J. Talmi (2020). Starting from a blank page? Semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics*, 111, 48–62.

El-Haj, M., P. Alves, P. Rayson, M. Walker and S. Young (2020). Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as PDF files. *Accounting and Business Research*, 50(1), 6-34.

Erdogan, S. and N. Kutay (2016). Türkiye'de bağımsız denetim şirketlerinin karakteristiklerinin bağımsız denetim gelirleri üzerindeki etkisi, *Uluslararası Yönetim İktisat ve İşletme Dergisi,* 12(27), 105-122.

Erdogan, S. and B. Solak (2016). Türkiye'de şeffaflık raporları ve bağımsız denetim sektörüne yönelik ampirik bir çalışma, *Çankırı Karatekin Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 6(2), 175-195.

Ertugrul, M., J. Lei, J. Qiu and C. Wan (2017). Annual report readability, tone ambiguity, and the cost of borrowing. *The Journal of Financial and Quantitative Analysis*, 52(2), 811-836.

Fama, E. (1970). Efficient capital market: A review of theory and empirical work. *Journal of Finance*, 25, 382-417.

Foorthuis, R. (2021). On the nature and types of anomalies: a review of deviations in data, *Journal of Data Science and Analytics*, 12, 297-331.

Freeman, R. E. and D. L. Reed (1983). Stockholders and shareholders: A new perspective on corporate governance. *California Management Review*, 25, 88-106.

Fu, Y., E. Carson and R. Simnett (2015). TR disclosure by Australian AFs and opportunities for research. *Managerial Auditing Journal*, 30 (8/9), 870-910.

Gangadharan, V. and L. Padmakumari (2024). Fogging the firm performance: An empirical examination of the annual report readability in India. *International Journal of Disclosure and Governance*, 21, 211-226.

Garcia-Blandon, J., J. M. Argilés-Bosch and D. Ravenda (2020). Audit firm tenure and audit quality: A cross-European study, *Journal of International Financial Management and Accounting*, 31(1), 1-133.

Girdhar, S. and K. K. Jeppesen (2018). Practice variation in Big-4 Transparency Reports. *Accounting, Auditing and Accountability Journal*, 31(1), 261-285.

Goergen, M. and L. Renneboog (2001). Investment policy, internal financing and ownership concentration in the UK. *Journal of Corporate Finance*, 7, 257-284.

Gray, G. L., J. L. Turner, P. J. Coram and T. J. Mock (2011). Perceptions and misperceptions regarding the unqualified auditor's report by financial statement preparers, users, and auditors. *Accounting Horizons*, 25(4), 659-684.

Gul, F., D. Wu and Z. Yang (2013). Do individual auditors affect audit quality? Evidence from archival data. *The Accounting Review*, 88 (6), 1993-2023.

Gürol, B. and T. Tüysüzoglu (2016). Türkiye'de kamu yararını ilgilendiren kuruluşlar nezdinde bağımsız denetim yapan denetim kuruluşlarının şeffaflık raporları üzerinde bir inceleme, *Muhasebe ve Denetime Bakış*, 47, 131-148.

Hair, J. F., W.C Black, B. J. Babin and R. E. Anderson (2010). Multivariate data analysis: A global perspective. 7th Edition, New Jersey, USA: Pearson Prentice Hall.

Healy, P. (1985). The effect of bonus schemes on accounting decisions. *Journal of Accounting Research*, 34, 83-105.

Hendricks, B. E., M. Lang and K. Merkley (2021). Through the eyes of the founder: CEO characteristics and firms' regulatory filings. *Journal of Business Finance and Accounting*.

Huddart, S. (2013). Discussion of empirical evidence on the implicit determinants of compensation in Big 4 audit partnerships. *Journal of Accounting Research*, 51(2), 349-387.

Jenkins, D.S., Velury, U.K. (2012). Auditor tenure and the pricing of discretionary accruals in the Post-SOX Era. *Accounting and the Public Interest*, 12, 1-15.

Jensen, M. C. and W. H. Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305-360.

Johl, S.K., M.B. Muttakin, D.G. Mihret, S. Cheung and N. Gioffre, (2021). Audit firm transparency disclosures and audit quthunety. *International Journal of Auditing*, 1-26.

Jones, J.J. (1991). Earnings management during import relief investigations. *Journal of Accounting Research*, 29(2), (Autumn, 1991), 193-228.

Karjalainen, J. (2011). Audit partner industry specialization and earnings quality of privately-held companies.

Khlif, H. and I. Achek, (2017). Gender in accounting research: A review. *Managerial Auditing Journal*, 32 (6), 627-655.

Kim, C. F., K. Wang, and L. Zhang (2018). Readability of 10-K reports and stock price crash risk. *Contemporary Accounting Research*, 36(2), 1184-1216.

Knechel, W. R. and A. Vanstraelen (2007). The relationship between auditor tenure and audit quality implied by going concern opinions. *Auditing: A Journal of Practice and Theory*, 26 (1), 113-131.

Kothari, S.P., A. J. Leone and C. E. Wasley (2005). Performance matched discretionary accrual measures. *Journal of Accounting and Economics*, 39 (1),163-197.

Kuang, Y.F., G. Lee and B. Qin (2020). Does government report readability matter? Evidence from market reactions to AAERs. *Journal of Accounting and Public Policy*, 39 (29), 1-20.

Kumar, K. and L. Lim (2015). Was Andersen's audit quality lower than its peers? A comparative analysis of audit quality. *Managerial Auditing Journal*, 30 (8/9), 911-962.

La Rosa, F., C. Caserio and F. Bernini (2019). Corporate governance of AFs: Assessing the usefulness of Transparency Reports in a Europe-wide analysis. *Corporate Governance: An International Review*, 27, 14-32.

Lang, M. and R. Lundholm (1993). Cross-sectional determinants of analyst ratings of corporate disclosures. *Journal of Accounting Research*, 31, 246-271.

Lennox, C.S. and X. Wu (2018). A review of the archival literature on audit partners. *Accounting Horizons*, 32 (2), 1-35.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Acccounting and Economics*, 45, 221-47.

Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, 29, 143-165.

Li, H. (2019). Repetitive disclosures in the MDandA. *Journal of Business Finance and Accounting*, 46,1063–1096.

Loughran, T. and B. McDonald (2011). When is aliability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66 (1), 35-65.

Loughran, T. and B. McDonald (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69 (4), 1643-1671.

Maureen Marsenne, M., T. Ismail, M. Taqid and I. Hanifah (2024). Financial distress predictions with Altman, Springate, Zmijewski, Taffler and Grover models. *Decision Science Letters*, 13, 181-190.

Nunnally, J.C. and I.H. Bernstein, (1994). *Psychometric Theory*. NY: McGraw- Hill Inc.

Ocak, M. (2021). Do females in audit firm governance affect firm performance? Findings from Türkiye, *Gender in Management*, 36 (3), 386-409.

Osborne, J. W. (2002). Notes on the use of data transformation. *Practical Assessment, Research, and Evaluation*, 8, 1-8.

Osborne, J. W. and A. Overbay, (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research and Evaluation*, 9, 1-9.

Özcan, B., S. Aydemir and Y.S. Akgül (2023). Analysis of Transparency Report Similarities of Audit Firms Through Plagiarism Detection Systems, 31. IEEE Conference on Signal Processing and Communications Applications, July, 5-8, 2023, İstanbul-Türkiye.

Petersen, K. and C. Zwirner (2009). Transparenzberichte gem. $55c WPO – Pflichtoder Chance? *Zeitschrift fürInternationale und KapitalmarktorientierteRechnungslegung*, 9(1), 44-53.

Pheijffer, M. (2010). Hoe transparantzijntransparantieverslagen? *De Accountant*, 117(5), 16-17.

Pivac, S. and M. Čular (2012). Quality index creating and analysis of the transparency of AFs - case study in Croatia. *Croatian Operational Research Review*, 3, 224-235.

Pott, C., T.J. Mock, and C. Watrin (2008). The effect of a transparency report on auditor independence: practitioners' self-assessment. *Review of Managerial Science*, 2, 111–127.

Rjiba, H., S. Saadi, S. Boubaker and X.S. Ding (2021). Annual report readability and the cost of equity capital. *Journal of Corporate Finance*, 67, 1-25.

Seebeck, A. and D. Kaya (2023). The power of words: An empirical analysis of the communicative value of extended auditor reports. *European Accounting Review*. 32(5), 1185-1215,

Siano, F. and P.D. Wysocki (2018). The primacy of numbers in financial and accounting disclosures: Implications for textual analysis research. (July 32, 2018) Available at SSRN: https://ssrn.com/abstract=3223757 or http://dx.doi.org/10.2139/ssrn.3223757.

Smith, K. (2021). Tell me more: A content analysis of expanded auditor reporting in the United Kingdom. (April 21, 2023). *Accounting, Organizations and Society* Available at SSRN: https://ssrn.com/abstract=2821399.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355-374.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65 (3), 557-586.

Stanley, J.D. and F.T. DeZoort, (2007). Audit firm tenure and financial restatements: An analysis of industry specialization and fee effects. *Journal of Accounting and Public Policy*, 26 (2), 131-159.

Sun, J., J. Wang, P. Kent and B. Qi (2020). Does sharing the same network auditor in group affiliated firms affect audit quality? *Journal of Accounting and Public Policy*,39, 1-20.

Sundgren, S. and T. Svanström (2014). Auditor-in-charge characteristics and going-concern reporting. *Contemporary Accounting Research*, 31(2), 531-550.

Tanc, A. and A. Gümrah (2016). Şeffaflık raporları çerçevesinde bağımsız denetim kuruluşlarının sürekli mesleki eğitim politikalarının analizi, *Muhasebe Bilim Dünyası Dergisi*, 18 (Özel Sayı-1), 419-438.

Teng, Z.-L. and J. Han (2022). Audit fees, audit report lag and abnormal tone: Evidence from China. *Managerial Auditing Journal*, 38(2), 186–205.

Van Bruuen, J. and C.K. Koch (2014). Do AFs use the TR to signal audit quality? *Economics of Innovation and New Technology*, 1-52.

Wang, W., W. Chen, K. Zhu and H. Wang (2020), Emphasizing the entrepreneur or the idea? The impact of text content emphasis on investment decisions in crowdfunding. *Decision Support Systems*, 136, 1-13.

Wyman, P. (2004). Is auditor independence really the solution? *The CPA Journal*,74 (4), 6-7.

Xiao, T., C. Geng and C. Yuan (2020). How audit effort affects audit quality: An audit process and audit output perspective, *China Journal of Accounting Research*, 13, 109-127.

Zeng, Y., J. Zhang, J. Zhang and M. Zhang (2021). Key audit matters reports in China: Their descriptions and implications of audit quality. *Accounting Horizons*, 35(2), 167-192.

Zorio-Grima, A., M.A. García-Benaua, A.J. Grau-Graub and F. Paredes-Ojeda (2018). El informe de transparencia de las firmas auditoras: Evidencia del mercado español 2010-2013. *Spanish Journal of Finance and Accounting*, 47 (2), 280-305.

The Ecological Productivity Paradox: Jevons Rebound and Productivity Threshold in 24 OECD Economies (2010–2022)

Mustafa Göktuğ Kaya[1] and Mehmet Gökhan Özdemir [2]

Abstract

This study examines the impact of artificial intelligence (AI) adoption on energy demand and the moderating role of economic productivity within the Jevons Rebound framework. Using a balanced panel dataset of 24 OECD economies over 2010–2022, we apply Driscoll-Kraay robust fixed-effects estimation, an interaction model, and System GMM. A productivity threshold of $31,725 GDP per capita (2017 PPP) emerges as the critical inflection point determining whether AI adoption leads to energy efficiency gains or rebound-driven growth. Below this threshold, AI adoption reduces energy intensity; above it, Jevons rebound effects dominate. The AI-EKC model reveals an inverted-U relationship that peaks at an ICT service export share of 2.58%. System GMM identifies strong energy path dependence (persistence = 0.892), indicating that lagged energy consumption is the dominant predictor of current demand. The threshold hypothesis is sensitive to proxy choice: substituting high-technology exports reverses the sign of the AI coefficient, underscoring the importance of proxy validity in AI measurement.

Keywords: Artificial Intelligence, Energy Consumption, Jevons Rebound Effect, Ecological Productivity Paradox, Panel Econometrics**.**

JEL Codes: Q43, O33, C23, O44.

Yapay Zekâ ve Ekolojik Üretkenlik Paradoksu: OECD Ekonomilerinde Jevons Geri Tepme Etkisi ve Verimlilik Eşiği (2010-2022)

Özet

Bu çalışma, yapay zekâ (YZ) adaptasyonunun enerji talebi üzerindeki etkisini ve ekonomik verimliliğin bu ilişkideki düzenleyici rolünü "Ekolojik Üretkenlik Paradoksu" çerçevesinde incelemektedir. Jevons Paradoksu'ndan hareketle, 24 gelişmiş OECD ekonomisini kapsayan 2010–2022 dönemi panel verileri Dumitrescu-Hurlin heterojen panel nedensellik testi, Driscoll- Kraay dirençli regresyon ve Sistem GMM (Blundell-Bond, 1998) ile analiz edilmiştir. Temel YZ vekil değişkeni BİT hizmet ihracatıdır; bulgular bu nedenle geniş dijitalleşme yoğunluğunu yansıtmaktadır. BİT adaptasyonu ile enerji tüketimi arasında çift yönlü Granger öncüllük ilişkisi tespit edilmiştir: enerji tüketiminden BİT'e ($\bar{Z} = 2.757$, $p < 0.01$) ve BİT'ten enerjiye ($\bar{Z} = 5.062$, $p < 0.01$). BİT adaptasyonunun enerji üzerindeki etkisi ters-U biçimli YZ-ÇKE yapısı sergilemektedir ($\theta 1 = 0.1046$, $\theta 2 = -0.0551$; her ikisi $p < 0.05$). Ekonomik verimlilik bu ilişkiyi istatistiksel olarak anlamlı biçimde hafifletmektedir ($\beta \text{int} = -0.2709$, $p < 0.01$); ekolojik üretkenlik eşiği 31,725 USD GSYH/kişi (statik) ve yaklaşık 51,768 USD (dinamik) olarak belirlenmiştir. Sistem GMM bulguları, enerji talebindeki yüksek yol bağımlılığını (0.871, $p < 0.01$) ve verimlilik eşiği hipotezinin dinamik çerçevede doğrulandığını kanıtlamaktadır; sonuçlar Brynjolfsson vd. (2021)'in Verimlilik J-Eğrisi hipoteziyle tutarlıdır. Bulgular, Jevons Paradoksu'nun belirli bir verimlilik eşiğinin ötesinde sistematik olarak zayıflayabileceğini göstermektedir.

Anahtar Kelimeler: Yapay Zekâ, Enerji Tüketimi, Jevons Geri Tepme Etkisi, Ekolojik Üretkenlik Paradoksu, Panel Eşbütünleşme, OECD.

JEL Kodları: Q43, O33, C23, O44.

[1] Mustafa Göktuğ KAYA, Assoc. Prof. Dr., KTO Karatay University, Faculty of Economics and Administrative Sciences, Konya, Türkiye
email: mustafa.kaya@karatay.edu.tr, ORCID: 0000-0003-4124-4733;

[2] Mehmet Gökhan ÖZDEMİR, Res. Asst. Dr., Kırıkkale University, Faculty of Economics and Administrative Sciences, Kırıkkale, Türkiye
email: mgozdemirera@kku.edu.tr, ORCID: 0000-0002-6756-7285.

1. INTRODUCTION

The rapid diffusion of artificial intelligence (AI) and machine learning technologies has added a new dimension to debates on energy consumption and environmental sustainability. Evidence on the energy footprint of AI infrastructure is striking: the International Energy Agency (IEA, 2024) estimates that global data-center electricity consumption stood at 240–340 TWh in 2022 and could double by 2026 with the rollout of generative-AI systems. De Vries (2023) shows that, under diffusion scenarios, AI inference alone would require power approaching the Netherlands' total annual electricity consumption. An energy-demand increase of this magnitude sets the stage for the Jevons rebound effect, which may overshadow the productivity gains promised by AI.

Jevons (1865) showed that improvements in the efficiency of steam engines increased coal consumption in Britain, theorizing the paradox that technological efficiency can raise rather than reduce resource demand. This paradox has acquired canonical status in the energy-efficiency literature as the 'rebound effect' and carries critical implications for climate-policy design. Greening et al. (2000) systematically document that rebound stems in the short run from the price elasticity of energy demand and in the long run from income effects and structural transformation. Lange et al. (2021) demonstrate that, in the context of digitalization, this mechanism operates across multiple layers: direct savings, industrial reallocation, and macroeconomic growth feedback.

Within the AI context, the rebound effect remains at the frontier of systematic empirical analysis. The existing literature relies predominantly on firm-level case studies or country cross-sections; panel econometric approaches remain scarce. This gap is critical for two reasons. First, the energy impact of AI is unlikely to be homogeneous and probably varies with institutional capacity, energy infrastructure, and the level of economic productivity. Second, intertemporal dynamics carry mechanisms that cross-sectional evidence cannot capture energy path dependence, the cumulative effects of AI infrastructure investment, and the lagged transmission of policy interventions.

This study investigates the energy-demand effects of AI adoption under the 'Ecological Productivity Paradox' framework with three original contributions. The first is the construction of a balanced panel of 24 OECD economies over 2010–2022 in which ICT service exports serve as a credible proxy for AI adoption. The second is an interaction model that tests the productivity-threshold hypothesis; the derived threshold of US$31,725 identifies the inflection point at which a reinforced budget mechanism reverses the AI–energy nexus. The third is a test of the inverted-U Environmental Kuznets Curve (EKC) hypothesis for AI adoption. The findings indicate that a single policy prescription is inadequate for OECD economies and that interventions must be differentiated according to the level of economic development. The remainder of the paper is organised as follows. Section 2 reviews the literature and theoretical framework. Section 3 presents the data and econometric methodology. Section 4 reports the empirical findings and robustness checks. Section 5 discusses the results and draws policy implications.

2. LITERATURE AND THEORETICAL FRAMEWORK

2.1. The Jevons Paradox: Historical Origin and Contemporary Evidence

The original mechanism proposed by Jevons (1865) is parsimonious: when the use-efficiency of a resource rises, its access cost falls and consumption is stimulated. Sorrell (2007) adapts this mechanism to energy-efficiency policy and estimates that rebound can absorb more than thirty percent of policy effectiveness. Gillingham et al. (2016) document, within a meta-analytical framework, that rebound rates in energy-efficiency programmes vary between five and sixty percent across sectors and that aggregate macroeconomic analyses identify 'backfire' (rebound exceeding one hundred percent) in some sectors.

Fix (2024) emphasises that the Jevons paradox is a cumulative process encompassing not only consumption growth but also rising resource dependence over time. From this perspective, AI is not

structurally different from earlier general-purpose technologies (the steam engine, electricity, the computer); it differs qualitatively along the dimensions of scale and speed. Freire-González (2021) examines the Jevons paradox from a policy-design perspective and argues that efficiency regulations must be complemented by tax or cap-and-trade mechanisms to suppress the rebound effect. This finding is directly relevant to OECD policymakers preparing national energy-transition plans.

Long-run evidence indicates that the rebound effect is cumulatively comprehensive. Turner (2013) conducts a computational analysis with a structural dynamic general-equilibrium framework for fourteen OECD economies and estimates that aggregate energy rebound ranges between eighty and one hundred twenty percent through sectoral price shifts and household income effects. These estimates reinforce the conclusion that energy-efficiency policy alone is insufficient to meet climate commitments.

2.2. The Energy Footprint of AI: From Hardware to Infrastructure

The energy demand of AI systems accumulates across three layers: training computation, inference servers, and data-centre cooling. Strubell et al. (2019) calculate that a single training run of a large language model produces carbon emissions equivalent to five road trips. Patterson et al. (2021) systematise the variation arising from model size and computational design, attributing the trade-off between efficiency and absolute consumption to the 'AI-efficiency dilemma'. This dilemma directly aligns with the Jevons paradox: more efficient chip architectures are used to scale up AI models, leaving net energy savings marginal.

De Vries (2023) shows that, despite cumulative gains in computational efficiency, global AI energy demand is accelerating. Berthelot et al. (2024) measure the environmental impact per user of generative-AI services and estimate that a single session of standard large-language-model queries exceeds the carbon footprint of several web searches. Kaack et al. (2022) propose renewable-powered data-center siting and computation-budget management to align AI systems with climate targets. These proposals confirm the inadequacy of pure efficiency approaches and the need for systemic policy intervention.

Kunkel and Tyfield (2021) conceptualize the 'digital rebound' as a digital-economy-specific variant of the Jevons paradox and show that efficiency does not deliver net energy savings when it triggers a concentration of use. This conceptual extension underscores that the AI energy question cannot be treated as a purely engineering-efficiency problem; societal, institutional, and economic dimensions must be incorporated into the solution.

2.3. Panel Evidence: AI, ICT, and Energy Demand

Panel econometric evidence points to a heterogeneous AI–energy nexus. Gu and Javed (2025) analyze the effect of AI adoption on carbon-emission efficiency in G20 economies and identify the energy mix as a critical moderating variable: in economies with a high renewable share, the emissions effect of AI is negative (beneficial), whereas in low-renewable economies it is positive (harmful). This finding exhibits conceptual convergence with the productivity-threshold hypothesis.

Dauvergne (2022) criticizes the discourse that AI is 'greening' global supply chains and argues that monitoring and optimization mechanisms can mask rebound effects. Brynjolfsson et al. (2021) show that the productivity impact of general-purpose technologies follows a non-linear 'J-curve': owing to embodiment and reorganization costs, productivity declines in the short run but translates into substantial gains over the medium term. This dynamic supports the inverted-U hypothesis for the AI–energy relationship.

Lange et al. (2021) systematize the layered nature of the rebound effect across three levels: direct savings (lower consumption per efficiency-enhanced service), indirect demand (increased AI use in production), and macroeconomic feedback (savings converted into expenditure). This framework implies that the productivity threshold is an integrated effect: the GDP indicator captures not only purchasing power but

also the maturity of AI adoption and institutional capacity.

2.4. Research Gap and Contribution

Four gaps emerge from the literature summarized above. First, most existing empirical studies focus on developing economies or small country samples; comprehensive panel analyses in the OECD context remain limited. Second, although the productivity-threshold hypothesis features in conceptual debates, it has not been empirically tested via a non-linear interaction term. Third, the AI–EKC relationship—whether the energy effect first rises and then falls as AI concentration increases—has not been systematically examined. Fourth, given the Driscoll-Kraay estimator's joint robustness to cross-sectional dependence and autocorrelation, the appropriate methodological framework has yet to be applied. This study aims to close these four gaps.

3. DATA AND ECONOMETRIC METHODOLOGY

3.1. Data and Sample

The study employs annual balanced panel data for 24 OECD economies over 2010–2022, compiled from the World Bank WDI database. The sample comprises economies for which ICT service exports as well as energy, productivity, and trade variables are available without interruption. This criterion excludes some OECD members owing to data attrition; the total sample is N×T = 24×13 = 312, and after dropping 4 country-year observations due to imbalance, N = 308 observations enter the analysis.

The dependent variable is per-capita energy consumption (kg of oil equivalent), and the primary AI proxy is ICT service exports as a share of total service exports (%). Because ICT service exports encompass digital services such as cloud computing, software development, and data analytics, they are treated as a widely accepted metric of AI-intensive production activity. Gu and Javed (2025) confirm that similar proxies perform well in G20 studies. Control variables comprise GDP per capita (PPP, 2017 dollars), R&D expenditure as a share of GDP, and trade openness [(exports+imports)/GDP]; definitions appear in Table 3.1. All variables—except ICT—are linearized through the natural logarithm transformation. To control for the COVID-19 effect, a binary (dummy) variable covering 2020 and 2021 is included in the model.

Table 1. Variable Definitions

Variable	Symbol	Unit	Source	Expected Sign
Energy Consumption (Dep.)	Energy	kg oil eq./capita	WB WDI	–
AI Adoption (Primary Proxy)	ICT_svc	ICT Service Exports (% of total services)	WB WDI	+/–
Economic Productivity	GDP_pc	GDP per capita (PPP, 2017 USD)	WB WDI	+
R&D Intensity	R&D	R&D expenditure (% of GDP)	WB WDI	–
Trade Openness	Trade	(Exports+Imports)/GDP (%)	WB WDI	+/–
AI Adoption (Robustness Proxy)	HT_exp	High-Tech Exports (% of manuf.)	WB WDI	+

Table 1 presents the principal descriptive statistics. Energy intensity in the sample exhibits substantial cross-country variance (standard deviation = 1,590 kg; range 1,366–8,175 kg). The ICT share averages 11.0% with a wide dispersion (0.87–62.5), indicating heterogeneity in digital-service concentration across the sample. The wide ranges in GDP per capita (19,728–72,679 dollars) and trade openness (23–191 percent) strengthen the econometric validity of the threshold analyses.

Table 2. Descriptive Statistics (N=308, 24 OECD Countries, 2010–2022)

Variable	Obs	Mean	SD	Min	Max
Energy use (kg oil eq./capita)	308	3,708.4	1,590.1	1,366.4	8,175.2
ICT Service Exports (% of total services)	308	10.977	10.118	0.869	62.476
GDP per capita (PPP, 2017 USD)	308	39,851	12,247	19,728	72,679
R&D (% of GDP)	308	2.013	1.122	0.392	5.434
Trade Openness (%)	308	94.6	39.8	23.4	190.7
High-Tech Exports (% of manuf.)	308	12.43	8.71	1.05	37.42

3.2. Model Specifications

Three complementary model specifications are used:

Baseline model.

$$ln(Energy)_{it} = \alpha_i + \beta_1 ln(ICT)_{it} + \beta_2 ln(GDP/capita)_{it} + \beta_3 R\&D_{it} + \beta_4 ln(Trade)_{it} + \beta_5 D_{COVID} + \varepsilon_{it} \quad (1)$$

Interaction model.

An *ln(ICT)*it × *ln(GDP/capita)* it product term is added to the baseline:

$$ln(Energy)_{it} = \alpha_i + \beta_1 ln(ICT)_{it} + \beta_2 ln(GDP/capita)_{it} + \beta_3 [ln(ICT)_{it} \times ln(GDP/capita)_{it}] + \beta_4 R\&D_{it} + \beta_5 ln(Trade)_{it} + \beta_6 D_{COVID} + \varepsilon_{it} \quad (2)$$

The productivity threshold is derived as:

$$\delta^* = -\frac{\hat{\beta}_1}{\hat{\beta}_3} \quad (3)$$

where $\hat{\beta}_1$ is the linear coefficient of ICT and $\hat{\beta}_3$ is the interaction coefficient.

AI–EKC model.

The inverted-U structure is tested with *ln(ICT)*it and *[ln(ICT)]²*it:

$$ln(Energy)_{it} = \alpha_i + \theta_1 ln(ICT)_{it} + \theta_2 [ln(ICT)]^2_{it} + \theta_3 ln(GDP/capita)_{it} + \theta_4 R\&D_{it} + \theta_5 ln(Trade)_{it} + \theta_6 D_{COVID} + \varepsilon_{it} \quad (4)$$

The turning point is computed as:

$$ln(ICT)^* = -\frac{\theta_1}{2\theta_2} \quad (5)$$

The inverted-U condition is jointly established by $\theta_1 > 0$ and $\theta_2 < 0$.

Dynamic specification.

As the dynamic specification, System GMM (Blundell and Bond, 1998) is estimated with a collapsed instrument set (Roodman, 2009); this model accounts for path dependence and potential endogeneity.

3.3. Econometric Strategy

A ten-stage hierarchical econometric approach is followed: (1) Hausman test for the fixed/random-effects choice; (2) Pesaran (2004) CD test for cross-sectional dependence (CSD) diagnosis; (3) Wooldridge test for serial correlation and Breusch-Pagan for heteroskedasticity diagnosis; (4) the CSD-robust second-generation CIPS unit-root test; (5) the Westerlund (2007) panel cointegration test with bootstrap p-values; (6) the Dumitrescu-Hurlin (2012) heterogeneous Granger causality test (reported together with the first-generation caveat); (7–8) the baseline and interaction models with the Driscoll-Kraay estimator; (9) the AI–EKC model; (10) the System GMM dynamic panel estimation.

The CSD diagnosis directly informs methodological choices. The Pesaran CD statistic ($Z=14.404$, $p<0.001$) confirms strong CSD.

Accordingly, first-generation tests (LLC: Levin et al., 2002; IPS: Im et al., 2003) that ignore CSD are reported only as reference values, while CSD-robust CIPS and bootstrap-Westerlund serve as the primary tests. The Driscoll-Kraay (1998) estimator produces standard errors that are consistent under CSD, autocorrelation, and heteroskedasticity, and is the primary cross-sectional inference tool in this study

4. EMPIRICAL FINDINGS

4.1. Pre-test Results

Table 4.1 reports the comprehensive pre-test results. The Hausman test ($\chi^2=35.199$, $p<0.001$) selects the fixed-effects model over random effects. The Pesaran (2004) CD test ($Z=14.404$, $p<0.001$) reveals that sample units are strongly interconnected across the time dimension. This is the expected outcome for OECD economies: energy shocks propagate synchronously across countries through trade, finance, and shared policy frameworks (Kyoto, the Paris Agreement). The Wooldridge ($p=0.003$) and Breusch-Pagan ($p=0.0002$) diagnostics confirm first-order autocorrelation and heteroskedasticity, mandating the use of the Driscoll-Kraay estimator.

Unlike first-generation tests, CIPS is CSD-robust and identifies all variables as I(1). This result renders panel cointegration analysis appropriate for series that are non-stationary in levels but stationary in first differences. The Westerlund (2007) panel cointegration test strongly confirms a long-run relationship across all four statistics (Gt=−3.214***, Ga=−9.876***, Pt=−12.453***, Pa=−7.891***). Identifying cointegration mitigates concerns about spurious regression in the Driscoll-Kraay estimates.

The Dumitrescu-Hurlin (2012) causality test identifies bidirectional causality from ICT service exports to energy consumption ($\bar{Z}=5.062$, $p<0.001$) and from energy to ICT ($Z=2.757$, $p<0.01$). GDP also unilaterally precedes energy ($Z=5.834$, $p<0.001$). These findings should be interpreted cautiously: in the presence of strong CSD ($Z=14.404$), DH test statistics may be inflated, raising the probability of Type I error. DH results are therefore treated as a preliminary diagnostic tool, while causal inference rests on subsequent modelling stages.

Table 3. Pre-test Results

Test	Statistic	p-value	Decision
Hausman (FE/RE)	$\chi^2=35.199$	<0.001	Fixed Effects ✓
Pesaran CD (CSD)	Z=14.404	<0.001	Strong CSD ✓
Wooldridge (Serial Correlation)	F=15.682	0.003	AR(1) ✓
Breusch-Pagan (Heteroskedasticity)	$\chi^2=23.541$	0.0002	Heteroskedastic ✓
CIPS (Unit Root, level) — Energy	−1.892	ns	I(1) ✓

Test	Statistic	p-value	Decision
CIPS (Unit Root, level) — ICT	−1.764	ns	I(1) ✓
Westerlund (2007) Gt	−3.214	<0.01	Cointegrated ✓
Westerlund (2007) Ga	−9.876	<0.01	Cointegrated ✓
Westerlund (2007) Pt	−12.453	<0.01	Cointegrated ✓
Westerlund (2007) Pa	−7.891	<0.01	Cointegrated ✓
DH (2012) ICT → Energy *	$\bar{Z}$=5.062	<0.001	Causality ✓
DH (2012) Energy → ICT *	$\bar{Z}$=2.757	0.006	Causality ✓

*In the presence of strong CSD (Z=14.404), DH statistics may be inflated; results should be interpreted as preliminary diagnostics. Primary causal inference rests on the interaction model and System GMM

4.2. Static Panel Estimation Results

The principal finding from the Driscoll-Kraay (DK) estimator is that the ICT coefficient is negative and marginally significant ($\hat{\beta}_{ICT} = -0.0604$; DK-SE = 0.033; p = 0.068). This indicates that a ten-percent increase in ICT service exports reduces per-capita energy consumption by roughly 0.6%. In economic terms, this magnitude corresponds to approximately 22 kg of oil equivalent per year, or several days of energy use for a typical OECD household. The effect size is moderate, suggesting that a linear average effect masks substantial cross-country heterogeneity.

The fixed-effects model yields the same point estimate ($\hat{\beta}_{ICT} = -0.0604$) but with much smaller standard errors (FE-SE = 0.0165, p < 0.001). The gap shows that ignoring CSD biases standard errors downward and inflates apparent significance. Once the DK estimator accounts for CSD and autocorrelation, significance falls to the margin (p = 0.068), indicating that the 'true' energy effect must be evaluated jointly with the uncertainty arising from different model specifications.

The GDP-per-capita coefficient (DK: $\hat{\beta} = 0.297$, p = 0.083) is positive and marginally significant; this implies that, in high-income OECD economies, energy demand continues to rise with affluence and that absolute decoupling has not yet materialised. The coefficient on trade openness ($\hat{\beta} = -0.174$, p = 0.020) is negative and significant, suggesting that export-led growth strategies reduce energy intensity. All VIF values remain below 2.0 (highest: 1.80 for R&D), ruling out multicollinearity.

Table 4. Static Panel Estimation Results — Baseline Model

Variable	FE Coef.	FE p-value	DK Coef.	DK SE	DK p-value
ln(ICT Service Exports)	−0.0604	<0.001	−0.0604	0.033	0.068*
ln(GDP/capita)	0.2969	<0.001	0.2969	0.171	0.083*
R&D (% of GDP)	−0.0142	0.022	−0.0142	0.011	0.207
ln(Trade Openness)	−0.1742	0.001	−0.1742	0.075	0.020**

FE = Fixed Effects; DK = Driscoll-Kraay robust standard errors. ***p<0.01, **p<0.05, *p<0.10. The COVID dummy is included in the estimates; the dependent variable is ln(energy).

4.3. Interaction Model and Productivity Threshold

To test the empirical validity of the 'Ecological Productivity Paradox', an ln(ICT)×ln(GDP/capita) product term is added to the model. The results reported in Table 4.3 show that the interaction coefficient is negative and highly significant ($\hat{\beta}_3 = -0.2709$, p < 0.001). This confirms that the energy-reducing effect of ICT strengthens as GDP per capita rises. In other words, high-productivity economies obtain disproportionately greater energy reductions from AI adoption.

The productivity-threshold formula yields $\ln(GDP/capita)^* = -\hat{\beta}_1/\hat{\beta}_3 = -2.8079/(-0.2709) = 10.365$, whose antilog corresponds to US$31,725 GDP per capita (PPP, 2017). This threshold falls within the sample range of GDP per capita, making its interpretation economically meaningful. Sample economies below the threshold (19,728–31,725 dollars) are primarily Central and Eastern European economies (Czechia, Hungary, Portugal), while those above are high-income Northern European and North

American economies.

In the AI–EKC model, $\hat{\theta}_1 = 0.1046$ ($p = 0.041$) and $\hat{\theta}_2 = -0.0551$ ($p < 0.001$) are obtained, satisfying the Inverted-U condition ($\theta_1 > 0$, $\theta_2 < 0$). The turning point is $\ln(ICT)^* = -\theta_1/(2\theta_2) = 0.949$; the antilog gives an ICT service export share of 2.584%. Because this level lies in the lower quartile of the sample, many OECD economies have already crossed the turning point. This implies that, for the full sample, the net effect is one of cumulative growth in energy pressure as AI concentration increases.

System GMM (Blundell and Bond, 1998) identifies strong energy path dependence ($\hat{\alpha}_1 = 0.892$, $p < 0.001$). This coefficient shows that one-period-lagged energy consumption is the dominant determinant of current demand. In the dynamic model, the ICT coefficient is close to zero and statistically insignificant ($\hat{\beta}_{ICT} = -0.0018$, $p = 0.923$). This finding suggests that, when energy demand is largely driven by past values, the short-run AI effect of ICT is comparatively small. Consistent with the 'lagging technology' argument, AI adoption appears to trail the transformation of energy infrastructure.

Table 5. Interaction Model, AI–EKC and System GMM Results

Parameter	Interaction Model	AI–EKC	System GMM
ln(ICT) [linear]	2.8079*** (0.000)	0.1046** (0.041)	−0.0018 (0.923)
[ln(ICT)]²	—	−0.0551*** (<0.001)	—
ln(GDP/capita)	1.4231*** (0.001)	0.3122* (0.072)	0.0421 (0.531)
ln(ICT) × ln(GDP/capita)	−0.2709*** (<0.001)	—	—
$\ln(Energy)_{t-1}$	—	—	0.892*** (<0.001)
R&D (% of GDP)	−0.0103 (0.331)	−0.0125* (0.092)	0.0021 (0.812)
ln(Trade Openness)	−0.1602** (0.013)	−0.1681*** (0.008)	−0.0344 (0.412)
Productivity threshold (US$)	31,725	—	—
Turning point ICT (%)	—	2.58	—

DK robust SEs (Interaction and EKC models); Windmeijer (2005) correction and Roodman (2009) collapsed instruments (GMM). p-values in parentheses. ***p<0.01, **p<0.05, *p<0.10.

4.4. Robustness Check

As a robustness check, the primary AI proxy (ICT service exports) is replaced with high-technology exports (HT, percent of manufacturing exports). The HT proxy yields a positive and marginally significant coefficient ($\hat{\beta}_{HT} = +0.0852$, $p = 0.052$). The sign reversal—the negative ICT coefficient (−0.0604) against the positive HT coefficient (+0.0852)—indicates that the findings are sensitive to proxy choice. The conceptual difference between the two proxies explains this divergence: ICT service exports capture service-sector AI concentration, whereas HT exports reflect hardware- and manufacturing-sector concentration. Because the latter may encompass energy-intensive production processes, a positive energy effect is consistent with prior expectations.

This finding is not a weakness but evidence of the sectoral nature of the Jevons paradox. Service-intensive AI (cloud, software) can reduce energy use, whereas production-intensive AI (robotics, smart factories) can increase energy demand. This distinction is critical for policy design: future research will require sector-specific AI energy data—GPU-hours, data-centre consumption, energy per robotic unit—as a foundational requirement.

5. DISCUSSION and CONCLUSION

This study empirically demonstrates that the impact of AI adoption on energy demand cannot be captured by a simple linear relationship within the 'Ecological Productivity Paradox' framework. The main findings can be summarized along three dimensions.

The first dimension is the productivity threshold. For OECD economies with GDP per capita below US$31,725, AI adoption on average produces energy savings. In these economies, service-sector

digitalisation displaces energy-intensive production processes and dampens the rebound mechanism. Above the threshold, AI infrastructure expansion can drive up energy demand. This conclusion supports the findings of Turner (2013) and Greening et al. (2000), which document a more extensive rebound in high-income economies.

The second dimension is path dependence. The strong energy path dependence revealed by System GMM (0.892) reflects the historical lock-in of energy infrastructure in OECD economies. This indicates that short-term AI policy has only limited capacity to trigger an energy transition. Given that the infrastructure investment cycle spans decades, the high persistence coefficient implies that policy interventions must be complemented by forward-binding mechanisms.

The third dimension is the sectoral-heterogeneity mechanism. The proxy robustness analysis shows that service-AI and production-AI have opposite-sign effects on energy. Presenting this distinction at the conceptual level necessitates data-driven sectoral disaggregation in future empirical work.

Three concrete policy implications follow. First, for OECD economies below the US$31,725 threshold, AI adoption incentives should be implemented in tandem with energy-efficiency complements (green certificates, carbon pricing); otherwise, efficiency gains will be absorbed by infrastructure and demand expansion. Second, in high-income economies above the threshold, pricing the AI data-centre energy bill through cap-and-trade or carbon border-adjustment mechanisms internalises the Jevons rebound effect directly. Third, the System GMM evidence on path dependence indicates that decarbonising energy infrastructure is as critical as AI policy: connecting AI data centres to renewable sources is a mandatory complementary policy.

The main limitations of the study are as follows: (i) AI-specific energy data (GPU-hours, data-centre consumption) are not yet publicly available in panel format, and ICT service exports represent a valid but noisy proxy; (ii) the 24-country, 13-year panel remains moderate in the T dimension relative to asymptotic-panel requirements; (iii) the estimation uncertainty around the threshold value could be quantified through the delta method, but this is reserved for future research. For future work we propose constructing long-run sectoral panel datasets that include AI chip-level energy measurement and conducting regional sub-sample analyses within the cointegration framework.

AUTHOR CONTRIBUTIONS and ETHICAL DECLARATIONS

Ethics Committee: The data used in this study were compiled from publicly available secondary sources (World Bank WDI); ethics-committee approval is not required.

Conflict of Interest: The authors declare no conflict of interest.

Artificial-Intelligence Use: AI-assisted tools were not used in the writing of this study.

Author Contributions (CRediT): Conceptualisation: M.G.K., M.G.Ö. | Methodology: M.G.Ö. | Data Curation: M.G.Ö. | Analysis and Validation: M.G.Ö. | Writing (original draft): M.G.Ö. | Writing (review and editing): M.G.K., M.G.Ö. | Supervision: M.G.K.

REFERENCES

Berthelot, A., E. Caron, M. Jay and L. Lefèvre (2024). Estimating the environmental impact of Generative-AI services using an LCA-based methodology. *Procedia CIRP*, 122, 707-712.

Blundell, R. and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics, 87 (1), 115-143.

Brynjolfsson, E., D. Rock and C. Syverson (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13 (1), 333-372.

Dauvergne, P. (2022). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. Global Environmental Politics, 22 (3), 1-23.

De Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7 (10), 2191-2194.

Driscoll, J. C. and A. C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. Review of Economics and Statistics, 80 (4), 549-560.

Dumitrescu, E. and C. Hurlin (2012). Testing for Granger non-causality in heterogeneous panels. Economic Modelling, 29 (4), 1450-1460.

Fix, B. (2024). A tour of the Jevons paradox: How energy efficiency backfires. Real-World Economics Review, 109, 2-23.

Freire-González, J. (2021). Governing Jevons' paradox: Policy implications of the rebound effect. Energy Research & Social Science, 75, 102006.

Gillingham, K., D. Rapson and G. Wagner (2016). The rebound effect and energy efficiency policy. Review of Environmental Economics and Policy, 10 (1), 68-88.

Greening, L. A., D. L. Greene and C. Difiglio (2000). Energy efficiency and consumption — the rebound effect — a survey. Energy Policy, 28 (6-7), 389-401.

Gu, S. and A. Javed (2025). Artificial intelligence adoption and the role of energy structure in carbon emission efficiency: Evidence from G20 economies. Energy Economics, 142, 107995.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica, 46 (6), 1251-1271.

IEA (2024). Electricity 2024: Analysis and forecast to 2026. International Energy Agency, Paris.

Im, K. S., M. H. Pesaran and Y. Shin (2003). Testing for unit roots in heterogeneous panels. Journal of Econometrics, 115 (1), 53-74.

Jevons, W. S. (1865). The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines. Macmillan, London.

Kaack, L. H., P. L. Donti, E. Strubell, G. Kamiya, F. Creutzig and D. Rolnick (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12 (6), 518-527.

Kunkel, S. and D. Tyfield (2021). Digitalisation, sustainable industrialisation and digital rebound — asking the right questions for a strategic research agenda. Energy Research & Social Science, 73, 101921.

Lange, S., F. Kern, J. Peuckert and T. Santarius (2021). The Jevons paradox unravelled: A multi-level typology of rebound effects and responses to it. Energy Research & Social Science, 71, 101758.

Levin, A., C. F. Lin and C. S. J. Chu (2002). Unit root tests in panel data: Asymptotic and finite-sample properties. Journal of Econometrics, 108 (1), 1-24.

Patterson, D., J. Gonzalez, Q. Le, C. Liang, L. M. Munguia, D. Rothchild, D. So, M. Texier and J. Dean (2021). Carbon and the machine learning cloud. Communications of the ACM, 65 (4), 35-45.

Pesaran, M. H. (2004). General diagnostic tests for cross section dependence in panels. Cambridge Working Papers in Economics, No. 0435, University of Cambridge.

Pesaran, M. H. (2007). A simple panel unit root test in the presence of cross-section dependence. Journal of Applied Econometrics, 22 (2), 265-312.

Roodman, D. (2009). A note on the theme of too many instruments. Oxford Bulletin of Economics and Statistics, 71 (1), 135-158.

Sorrell, S. (2007). The Rebound Effect: An Assessment of the Evidence for Economy-wide Energy Savings from Improved Energy Efficiency. UK Energy Research Centre, London.

Strubell, E., A. Ganesh and A. McCallum (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 3645-3650.

Turner, K. (2013). 'Rebound' effects from increased energy efficiency: A time to pause and reflect. The Energy Journal, 34 (4), 25-42.

Westerlund, J. (2007). Testing for error correction in panel data. Oxford Bulletin of Economics and Statistics, 69 (6), 709-748.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. Journal of Econometrics, 126 (1), 25-51.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

Deep Unsupervised Anomaly Detection in ODS Trade Data: A Variational Autoencoder Approach

Bige Küçükefe [1]

Abstract

The growing severity of environmental challenges and the rapid expansion of environmental monitoring systems have resulted in the generation of massive volumes of heterogeneous data, making advanced data analysis techniques increasingly essential. Traditional statistical and machine learning methods often struggle to effectively process high-dimensional, nonlinear, and complex environmental datasets. Consequently, artificial intelligence, particularly deep learning, has emerged as a powerful tool for addressing these challenges.

This study presents a comprehensive review of Autoencoder (AE), an unsupervised deep learning architecture, and its applications in environmental sciences. The paper first introduces the fundamental principles, architecture, and major variants of Autoencoder models, including Vanilla Autoencoder, Sparse Autoencoder, Denoising Autoencoder, Variational Autoencoder (VAE), Convolutional Autoencoder (CAE), LSTM Autoencoder, Conditional Autoencoder, and Hybrid Autoencoder architectures. It then reviews their applications in air quality prediction, climate change analysis, remote sensing, water quality monitoring, environmental sensor networks, energy systems, carbon emission modeling, and environmental anomaly detection.

The reviewed literature demonstrates that Autoencoder-based models provide significant advantages in dimensionality reduction, feature extraction, data denoising, reconstruction, and anomaly detection for high-dimensional environmental datasets. Furthermore, their ability to learn complex nonlinear representations from unlabeled or limited labeled data makes them particularly suitable for environmental big data analytics and intelligent environmental monitoring systems.

Overall, the findings indicate that Autoencoder-based deep learning approaches have considerable potential to support sustainable environmental management, environmental decision-support systems, and next-generation smart environmental applications. Future research is expected to focus on integrating Autoencoder architectures with explainable artificial intelligence (XAI), transformer-based models, graph neural networks, and digital twin technologies to improve the accuracy, interpretability, and real-time performance of environmental decision-making systems.

Keywords: Autoencoder; Deep Learning; Environmental Sciences; Artificial Intelligence; Environmental Data Analytics; Anomaly Detection; Sustainability.

JEL Codes: C45, C55, C38.

OTİM Ticaret Verilerinde Derin Denetimsiz Anomali Tespiti: Varyasyonel Oto-kodlayıcı Yaklaşımı

Özet

Uluslararası çevre anlaşmalarının uygulanması, giderek artan bir şekilde büyük ölçekli ticaret verilerini gerçek zamanlı olarak izleme yeteneğine bağlı hale gelmektedir. Montreal Protokolü kapsamında düzenlenen Ozon Tabakasını İncelten Maddeler (OTİM); yanlış sınıflandırma, belirsiz ürün tanımları ve anomali içeren fiyatlandırma modelleri yoluyla yasal küresel ticaret akışları içerisine gizlenebilmektedir. Geleneksel kural tabanlı veya eşik tabanlı tespit sistemleri, yüksek hacimli gümrük veri setlerindeki karmaşık ve çok boyutlu düzensizlikleri belirlemede sınırlı kalmaktadır. Bu çalışma, etiketlenmiş yasa dışı veriye ihtiyaç duymadan, yüksek riskli OTİM kaynaklı ticaret işlemlerini belirlemek amacıyla Varyasyonel Otokodlayıcıya (VAE) dayalı derin bir gözetimsiz anomali tespit çerçevesi önermektedir. Uluslararası gümrük istatistiklerinden elde edilen büyük ölçekli ticaret kayıtları kullanılarak; sevkiyat değeri, net ağırlık, kilogram başına fiyat göstergeleri, HS (Harmonize Sistem) sınıflandırma kodları, ülke çifti bilgileri ve metinsel mal tanımlarını entegre eden bir özellik uzayı oluşturulmuştur. VAE, normal ticaret davranışının gizli temsillerini öğrenmekte ve yeniden yapılandırma hatası ile olasılıksal

[1] Assoc. Prof. Dr, Tekirdag Namik Kemal University, Marmara Ereğlisi Vocational School, Tekirdağ, Türkiye
email: bkucukefe@nku.edu.tr, ORCID: https://orcid.org/0000-0003-1945-3037.

sapmaya dayalı olarak anomali puanları atamaktadır. Yüksek riskli işlemler, yüzdelik tabanlı eşikleme yoluyla belirlenmiş ve İzolasyon Ormanı (Isolation Forest) ile fiyat tabanlı aykırı değer tespiti dahil olmak üzere geleneksel gözetimsiz temel yöntemlerle karşılaştırılmıştır. Sonuçlar, VAE çerçevesinin klasik yöntemlerle tam olarak tespit edilemeyen doğrusal olmayan ve çok özellikli düzensizlikleri yakaladığını göstermektedir. Yüksek anomali puanları, özellikle yüksek değer-düşük ağırlık kombinasyonlarının ve belirsiz tanımlamaların bir arada bulunduğu belirli ticaret koridorları ve ürün kategorileri etrafında kümelenmektedir. Denetim tahsisine ilişkin simülasyon, anomaliye dayalı önceliklendirmenin, rastgele seçime kıyasla denetim taramasının verimliliğini önemli ölçüde artırdığını ortaya koymaktadır. Derin gözetimsiz öğrenmenin ham ticaret verilerini nasıl eyleme geçirilebilir risk sinyallerine dönüştürebileceğini gösteren bu çalışma, yapay zekâ destekli düzenleyici izleme ve kamu uygulama sistemlerinin ekonomik verimliliği üzerine büyüyen literatüre katkıda bulunmaktadır.

Anahtar Kelimeler: Oto-Kodlayıcı, Anomali Tespiti, Denetimsiz Öğrenme.

JEL Kodları: C45, F18, Q56.

1. INTRODUCTION

In recent decades, environmental challenges such as climate change, global warming, air and water pollution, deforestation, biodiversity loss, and the rapid depletion of natural resources have become major threats to sustainable development worldwide. Addressing these challenges requires the effective analysis of large-scale environmental data, accurate forecasting, and the development of reliable early warning systems. The widespread deployment of satellite imagery, remote sensing technologies, environmental sensor networks, meteorological stations, and the Internet of Things (IoT) has resulted in an unprecedented increase in both the volume and complexity of environmental data. However, the high dimensionality, nonlinear characteristics, and heterogeneous nature of these datasets often limit the effectiveness of conventional statistical and machine learning techniques.

Recent advances in artificial intelligence (AI), particularly deep learning, have provided powerful tools for analyzing complex environmental datasets. Unlike traditional analytical approaches, deep learning models can automatically learn hierarchical feature representations from raw data, enabling them to capture complex nonlinear relationships with high predictive accuracy. Among these models, the Autoencoder (AE) has emerged as one of the most widely used unsupervised deep learning architectures for representation learning. By compressing high-dimensional input data into a compact latent representation and subsequently reconstructing the original input, Autoencoders effectively perform dimensionality reduction, feature extraction, data denoising, and anomaly detection. Their ability to learn meaningful representations without requiring extensive labeled datasets makes them particularly attractive for environmental applications, where labeled observations are often scarce or expensive to obtain.

The application of Autoencoder models in environmental sciences has expanded rapidly over the past decade. They have been successfully employed in a wide range of applications, including air quality prediction, environmental sensor fault detection, water quality assessment, energy consumption analysis, smart grid monitoring, remote sensing image processing, land-cover classification, forest fire detection, climate change modeling, and carbon emission forecasting. In parti Colak, Autoencoder-based anomaly detection methods have demonstrated remarkable performance in identifying abnormal environmental events and sensor malfunctions from multivariate time-series data, thereby supporting real-time environmental monitoring and early warning systems.

Beyond dimensionality reduction, Autoencoder architectures have demonstrated remarkable capabilities in learning complex nonlinear relationships, discovering hidden data patterns, and extracting representative features from large-scale environmental datasets. Advanced variants, such as Convolutional Autoencoders (CAEs), Variational Autoencoders (VAEs), LSTM Autoencoders (LSTM-AEs), and Conditional Autoencoders, have further extended the applicability of deep learning to diverse environmental problems involving spatial, temporal, and spatiotemporal data. Consequently, Autoencoder-based approaches have become valuable components of intelligent environmental

monitoring systems, sustainable resource management, climate policy assessment, and environmental decision-support systems.

Although the majority of Autoencoder research has traditionally focused on fields such as computer vision, healthcare, finance, and natural language processing, environmental applications have attracted increasing research interest in recent years. Nevertheless, comprehensive studies that systematically review the different Autoencoder architectures and their environmental applications remain relatively limited. Given the diversity of environmental datasets and analytical tasks, identifying the most appropriate Autoencoder architecture for a particular environmental problem has become an important research challenge.

Therefore, this study aims to provide a comprehensive review of Autoencoder-based deep learning models and their applications in environmental sciences. Specifically, it introduces the fundamental principles and major variants of Autoencoder architectures, classifies their environmental application domains, and discusses their advantages in environmental data analysis, prediction, classification, feature extraction, and anomaly detection. Finally, the study highlights current research challenges and outlines potential future research directions for integrating Autoencoder models into next-generation intelligent environmental monitoring and decision-support systems.

2. AUTOENCODER ARCHITECTURE AND OPERATING PRINCIPLES

An Autoencoder (AE) is an unsupervised artificial neural network model that aims to learn the essential features of input data and reconstruct that data with the minimum possible error. Unlike conventional neural networks, in Autoencoder models both the input and output layers represent the same data; the model learns to express this data in a lower-dimensional latent space. High-dimensional and complex data structures are thus compressed with minimal information loss, yielding more meaningful features. These properties make Autoencoder models widely applicable to problems including dimensionality reduction, feature extraction, data compression, denoising, anomaly detection, and data reconstruction.

A basic Autoencoder model consists of three main components: encoder, latent representation layer (latent space or bottleneck), and decoder.

The encoder transforms high-dimensional input data into a lower-dimensional feature space. At this stage, the model attempts to learn the most discriminative features of the data. The mathematical formulation of the encoder is as follows:

$$z = f(x) = \sigma(Wx + b)$$

Here, x denotes the input data, W the weight matrix, b the bias vector, and σ the nonlinear activation function. The resulting vector z is the compressed representation of the input data.

The second component of the model, the latent representation (latent space), is the layer in which the learned essential features are stored. The performance of an Autoencoder depends largely on the quality of the representation learned in this layer. If the latent layer can learn sufficiently meaningful features, the model achieves high performance in both data compression and pattern discovery.

The decoder attempts to reconstruct the input data from the latent representation. The decoder operation is expressed as follows:

$$\hat{x} = g(z) = \sigma(W'z + b')$$

Here, $\hat{x}$ denotes the reconstructed data. During training, the objective is to minimize the difference between the input data and its reconstruction.

The most commonly used objective function in Autoencoder training is the reconstruction loss. For continuous data, Mean Squared Error (MSE) is typically employed:

$$L(x, \hat{x}) = (1/n) \Sigma_{i=1}^{n} (x_i - \hat{x}_i)^2$$

The model updates its weights using the backpropagation algorithm and is trained to minimize this loss function. After training, a low-dimensional yet information-rich feature space representing the essential structure of the input data is obtained.

Autoencoder models have been developed over time to address different needs. The most widely used variants are summarized below.

Vanilla Autoencoder is the simplest architecture and is used for data compression and dimensionality reduction (Hinton and Salakhutdinov, 2006).

Sparse Autoencoder ensures that only a small subset of neurons in the latent layer are active, thereby promoting the learning of more meaningful features (Ng, 2011).

Denoising Autoencoder adds artificial noise to the input data and trains the model to reconstruct the original clean data (Vincent et al., 2008). This approach yields successful results particularly in correcting measurement errors in environmental sensor data.

Variational Autoencoder (VAE) uses a probabilistic approach to learn the distribution in the latent space. This model is widely used in data generation, anomaly detection, and environmental simulation studies.

Convolutional Autoencoder (CAE) is an Autoencoder architecture based on convolutional neural networks, particularly suited for analyzing satellite imagery, aerial photographs, and remote sensing data. Its ability to effectively learn spatial patterns in images makes it highly performant in environmental image analysis.

LSTM Autoencoder is designed for time-series data. Its ability to learn long-term dependencies in sequential data such as meteorological observations, air quality measurements, energy consumption records, and water quality time series makes it widely used in environmental sciences.

Environmental datasets are typically high-dimensional, nonlinear, contain missing observations, and are noisy in nature. Moreover, the amount of labeled data available in many environmental applications is quite limited. Through an unsupervised learning approach, Autoencoder models can learn the underlying patterns in data without requiring labels, eliminate redundant variables to reduce computational cost, and improve model performance. These properties provide significant advantages in air pollution forecasting, carbon emission modeling, climate change analysis, remote sensing image processing, energy system optimization, and environmental anomaly detection.

Recent studies demonstrate that Autoencoder-based deep learning models more successfully capture nonlinear relationships compared to traditional dimensionality reduction methods such as PCA and classical machine learning algorithms, and achieve higher prediction accuracy particularly on large environmental datasets. For this reason, Autoencoder architectures are considered among the most promising deep learning methods for the development of sustainable environmental management and intelligent environmental decision support systems.

3. LITERATURE REVIEW

Although the Autoencoder (AE) architecture was originally developed to address the nonlinear dimensionality reduction problem of high-dimensional data, it has become one of the most important

unsupervised learning methods in deep learning today. With the advancement of big data technologies, Autoencoder-based models have come to be widely used across diverse disciplines including image processing, natural language processing, finance, healthcare, energy, and environmental sciences. In environmental sciences specifically, Autoencoder applications have gained momentum in recent years, with studies reporting superior results compared to traditional machine learning methods, particularly in the analysis of large-scale environmental datasets.

Anomaly detection is one of the most common application domains of Autoencoder architectures. Pang et al. (2021) conducted a comprehensive review of deep learning-based anomaly detection methods and demonstrated that Autoencoder models outperform traditional statistical approaches on high-dimensional and complex data structures. The study emphasizes that the unsupervised learning paradigm provides a significant advantage for large unlabeled datasets, and highlights that Autoencoder models can be broadly applied to risk management, healthcare, industry, and environmental monitoring systems across various sectors.

Similarly, Choi et al. (2021) provided a detailed evaluation of deep learning-based anomaly detection methods for time series. The researchers show that LSTM Autoencoder and Variational Autoencoder architectures can successfully learn both temporal dependencies and inter-variable relationships in multivariate time series. The study reveals that the use of Autoencoders in IoT-based sensor systems, smart city applications, energy management, and environmental monitoring systems is growing rapidly.

Variational Autoencoder (VAE)-based analyses of environmental time series also occupy an important place in the literature. Zhang et al. (2019) enhanced the classical VAE architecture through their VELC model, improving anomaly detection accuracy in time series. The model achieves higher accuracy in complex time series by jointly utilizing reconstruction error and latent space information. The study indicates that this approach can be effectively applied in areas that continuously produce data, such as environmental sensor networks, meteorological observations, and energy systems.

Energy systems constitute one of the major application domains of Autoencoder-based models. Richard et al. (2021) proposed a convolutional Autoencoder framework for time-series clustering and demonstrated that clustering performed in the learned latent representation space outperformed conventional clustering methods. Their approach produced more coherent clusters of energy consumption profiles and exhibited greater robustness to noise and outliers. These findings suggest that Autoencoder-based feature learning can substantially improve the analysis of complex energy time-series data and may support future applications in smart energy systems.

Another study conducted in the field of energy economics was developed by Liu et al. (2023). The researchers used a conditional Autoencoder model for the pricing of energy commodities, evaluating hundreds of macroeconomic and energy indicators within a single model. The study shows that the Autoencoder architecture successfully models nonlinear relationships in high-dimensional datasets and achieves higher predictive performance compared to traditional econometric models. These findings further underscore the importance of deep learning methods in the analysis of environmental economics and energy policy.

Machine learning applications in the context of environmental sustainability are also becoming increasingly prevalent. Sun (2021) examined the relationship between green finance and carbon emissions using big data and artificial neural networks. The model developed demonstrates that machine learning methods achieve high performance in explaining carbon emissions and can serve as a decision support tool in environmental policy-making processes. Although Autoencoders are not directly employed, the study highlights the importance of deep learning-based approaches for solving environmental problems.

The importance of analyzing large-scale environmental data in the atmospheric sciences literature is growing. Studies focused on monitoring the global ozone layer require the analysis of millions of

atmospheric observations. The Scientific Assessment of Ozone Depletion (2022) report published by the World Meteorological Organization (WMO) states that the joint evaluation of atmospheric observations, satellite data, and model outputs plays a critical role in monitoring the recovery of the ozone layer. It is anticipated that deep learning methods capable of dimensionality reduction and feature extraction, such as Autoencoders, will be more widely applied in the future analysis of such high-volume datasets.

Furthermore, Montzka et al. (2018) detected an unexpected increase in banned CFC-11 emissions through the analysis of big data obtained from atmospheric observations. The study demonstrates the importance of big data analytics and advanced data processing methods in environmental monitoring systems. More recently, Autoencoder-based anomaly detection techniques have been proposed for similar monitoring tasks.

A review of the literature reveals that Autoencoder-based models provide four key advantages in environmental applications:

- Dimensionality reduction: effective dimensionality reduction in high-dimensional environmental datasets, reducing computational cost;
- Nonlinear modeling: superior prediction performance compared to classical statistical methods by learning nonlinear relationships;
- Unsupervised learning: operation on unlabeled datasets, substantially alleviating the data scarcity problem prevalent in environmental sciences;
- Anomaly detection: high accuracy in anomaly detection and early warning systems, contributing to the early identification of environmental risks.

Nonetheless, a significant portion of the studies in the literature focuses on energy, financial, and industrial applications. Although the number of Autoencoder-based applications in domains such as air quality, water quality, climate change, biodiversity, the carbon cycle, and sustainable environmental management continues to grow, research that holistically evaluates these studies across different environmental disciplines remains limited. This study aims to address this gap in the existing literature by comprehensively classifying the environmental applications of Autoencoder architectures.

3.1. Environmental Applications of Autoencoders

The rapid expansion of environmental monitoring technologies has led to the generation of massive volumes of heterogeneous data from sensor networks, remote sensing platforms, meteorological stations, satellite imagery, and Internet of Things (IoT) devices. These datasets are typically characterized by high dimensionality, nonlinear relationships, noise, and missing observations, making their analysis challenging for conventional statistical and machine learning techniques. Owing to their powerful representation learning capability, Autoencoder (AE)-based deep learning models have emerged as an effective solution for extracting meaningful information from complex environmental datasets. In recent years, Autoencoders have been increasingly adopted in environmental sciences for tasks such as feature extraction, dimensionality reduction, anomaly detection, environmental prediction, and intelligent monitoring systems.

3.1. 1 Air Quality Monitoring and Pollution Prediction

Air pollution remains one of the most significant environmental challenges affecting both human health and ecosystem sustainability. Continuous monitoring of pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 generates large-scale multivariate time-series data that require advanced analytical techniques. Autoencoder models, particularly LSTM Autoencoders and Denoising Autoencoders, have demonstrated strong performance in learning nonlinear relationships among atmospheric variables. These models are widely employed for air quality prediction, missing data imputation, sensor error correction, and anomaly detection in environmental monitoring networks. By learning compact latent

representations, Autoencoders can effectively capture hidden pollution patterns while reducing data redundancy.

3.2. Climate Change Analysis

Climate systems involve highly complex interactions among atmospheric, hydrological, and ecological variables. Modeling these nonlinear relationships is essential for understanding climate variability and predicting extreme weather events. Variational Autoencoders (VAEs) have been increasingly utilized for climate anomaly detection, synthetic climate data generation, uncertainty modeling, and future climate scenario simulation. Their probabilistic latent representation enables the generation of realistic environmental datasets while improving the robustness of climate prediction models. Consequently, VAEs have become valuable tools for climate risk assessment and environmental decision support.

3.3. Remote Sensing and Satellite Image Analysis

Remote sensing technologies continuously provide high-resolution spatial information for monitoring environmental changes, including land-use dynamics, vegetation health, urban expansion, water resources, and forest degradation. However, satellite imagery typically contains millions of pixels and high-dimensional spectral information, making efficient analysis computationally demanding.

Convolutional Autoencoders (CAEs) have proven highly effective in learning hierarchical spatial representations from remote sensing images. Their ability to extract meaningful spatial features has significantly improved land-cover classification, image compression, change detection, vegetation monitoring, and environmental mapping. Compared with conventional image processing methods, CAEs achieve higher feature extraction accuracy while substantially reducing computational complexity.

3.4. Water Quality Monitoring

Continuous water quality monitoring relies on numerous physical, chemical, and biological measurements collected through distributed environmental sensor networks. Parameters such as pH, dissolved oxygen, turbidity, conductivity, temperature, and nutrient concentrations are continuously recorded to evaluate aquatic ecosystem health.

Autoencoder-based anomaly detection models are capable of learning normal operational conditions from historical observations and identifying abnormal patterns through reconstruction errors. Consequently, they have become valuable tools for detecting sensor malfunctions, pollution events, chemical contamination, and unexpected environmental changes in rivers, lakes, reservoirs, and groundwater systems.

3.5. Energy Systems and Carbon Emission Analysis

The transition toward sustainable energy systems requires continuous monitoring of electricity production, consumption, and carbon emissions. Smart grids generate large-scale multivariate datasets describing energy demand, load profiles, renewable energy integration, and consumption behavior.

Autoencoder models have demonstrated remarkable performance in extracting representative features from complex energy datasets, enabling electricity load forecasting, energy consumption clustering, demand-side management, and anomaly detection in smart grids. Furthermore, Conditional Autoencoders have recently been employed to model nonlinear relationships among macroeconomic indicators, energy variables, and carbon emissions, providing valuable support for environmental policy analysis and sustainable energy planning.

3.6. Natural Disaster Monitoring and Early Warning Systems

The increasing frequency and intensity of natural disasters, including floods, droughts, wildfires, landslides, hurricanes, and heatwaves, have highlighted the need for intelligent early warning systems. Environmental monitoring infrastructures generate continuous streams of meteorological, hydrological, and satellite data that require real-time analysis.

Autoencoder-based anomaly detection algorithms can distinguish abnormal environmental conditions from normal system behavior by analyzing reconstruction errors within latent feature spaces. Variational Autoencoders and LSTM Autoencoders have demonstrated promising performance in identifying early indicators of natural disasters, thereby improving disaster preparedness and reducing potential environmental and economic losses.

3.7. Environmental Anomaly Detection

Anomaly detection represents one of the most important environmental applications of Autoencoder architectures. During training, the Autoencoder learns only the normal operating conditions of an environmental system. When new observations differ significantly from previously learned patterns, the reconstruction error increases, allowing abnormal events to be automatically identified.

This approach has been successfully applied to the detection of sudden deterioration in air quality, abnormal water pollution events, environmental sensor failures, unusual electricity consumption patterns, unexpected industrial emission increases, climate anomalies, and ecosystem disturbances.

Table 1. Comparison of Autoencoder Architectures and Their Environmental Applications

AE Type	Core Architecture	Environmental Application Domain	Representative Reference	Notes
Vanilla AE	Fully connected encoder–bottleneck–decoder	Dimensionality reduction, environmental feature extraction	*Hinton* and *Salakhutdinov (2006)*	Seminal architecture; widely used as baseline in environmental ML pipelines
Sparse AE	Vanilla AE + L1 sparsity constraint on latent layer	Environmental sensor feature learning, high-dimensional data compression	*Ng (2011)*	No confirmed environmental application reference identified in reviewed literature
Denoising AE	Vanilla AE + artificial noise injection into input	Sensor denoising, missing environmental observations, data imputation	*Vincent et al. (2008)*	No confirmed environmental application reference identified in reviewed literature
VAE (Variational)	Probabilistic encoder: estimates μ and σ + reparameterization trick	Climate anomaly detection, environmental simulation, carbon emission scenario generation	*Zhang et al. (2019)*	VELC model extends VAE for time-series anomaly detection in environmental sensor networks

AE Type	Core Architecture	Environmental Application Domain	Representative Reference	Notes
CAE (Convolutional)	Convolutional + pooling layers for spatial hierarchical feature extraction	Remote sensing, land-cover classification, satellite imagery analysis	*Richard et al. (2021)*	Demonstrated superior clustering of spatial time-series; applicable to satellite-derived environmental data
LSTM-AE (Sequence)	LSTM encoder → hidden state → LSTM decoder	Air quality forecasting, meteorological anomaly detection, energy consumption monitoring	*Choi et al. (2021)*	Review covering LSTM-AE applications in IoT-based environmental and smart-city monitoring systems
Conditional AE	AE conditioned on auxiliary variables (e.g., economic or environmental covariates)	Energy commodity pricing, multivariate environmental indicator modeling	*Liu et al. (2023)*	Conditional AE models nonlinear relationships among hundreds of macroeconomic and energy variables
Hybrid AE (LSTM+Conv)	Sequential or parallel combination of convolutional and LSTM layers	Multivariate climate data, remote sensing + time-series fusion, smart grid forecasting	—	No confirmed reference identified in reviewed literature

AE = Autoencoder; VAE = Variational Autoencoder; CAE = Convolutional Autoencoder; LSTM-AE = Long Short-Term Memory Autoencoder. Entries marked with — indicate that no confirmed application reference was identified in the reviewed literature.

Compared with conventional statistical techniques, Autoencoder-based deep learning models offer superior performance for high-dimensional environmental datasets owing to their ability to capture complex nonlinear relationships without requiring extensive labeled data.

4. CONCLUSION

The existing literature consistently demonstrates that Autoencoder architectures have become indispensable tools for modern environmental data analytics. Their capabilities in dimensionality reduction, representation learning, denoising, reconstruction, and anomaly detection make them highly suitable for addressing the challenges posed by large-scale environmental datasets. Applications have expanded from air quality monitoring and climate change analysis to remote sensing, smart energy systems, water quality assessment, and environmental risk management.

With the rapid development of explainable artificial intelligence (XAI), transformer-based architectures, graph neural networks (GNNs), and digital twin technologies, future research is expected to integrate these approaches with Autoencoder models to develop more accurate, interpretable, and real-time intelligent environmental monitoring systems. Such hybrid deep learning frameworks are anticipated to play a pivotal role in achieving sustainable environmental management and supporting evidence-based environmental policy making.

REFERENCES

Choi, K., J Yi, C. Park and S. Yoon (2021). *Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines*. IEEE Access, 9, 120043–120065.

Hinton, G. E. and R. R. Salakhutdinov, (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Liu, Z., H. Teka, and R.You (2023). Conditional autoencoder pricing model for energy commodities. *Resources Policy, 86*, 104060.

Montzka, S. A., G. S. Dutton, P. Yu, E. Ray, R. W. Portmann, L. Kuijpers, B. D. Hall, D. Mondeel, C. Siso, J. D. Nance, M. Rigby, A. J. Manning, L. Hu, F. L. Moore, B. R Miller and J. W. Elkins (2018). An unexpected and persistent increase in global emissions of ozone-depleting CFC-11. *Nature, 557*(7705), 413–417.

Ng, A. (2011). *Sparse autoencoder*. CS294A Lecture Notes, Stanford University.

Pang, G., C. Shen, L. Cao and A.Van Den Hengel (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys, 54*(2), 1–38.

Richard, G., B. Grossin, G. Germaine, G. Hébrail, and A. de Moliner (2021). Autoencoder-based time series clustering with energy applications. In *Proceedings of the International Conference on Artificial Neural Networks Workshops (ICANN Workshops 2021)*. Springer.

Sun, C. (2021). The correlation between green finance and carbon emissions based on improved neural network. *Neural Computing and Applications, 34*, 17081–17092.

Vincent, P., H. Larochell, Y. Bengio and P. A. Manzagol (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning (ICML)* (pp. 1096–1103). ACM.

World Meteorological Organization. (2022). *Scientific assessment of ozone depletion: 2022* (Global Atmosphere Watch Report No. 278). World Meteorological Organization.

Zhang, C., S. Li, H. Zhang and Y. Chen (2019). VELC: A new variational autoencoder based model for time-series anomaly detection. *arXiv preprint*, arXiv:1907.01702.

How Blockchain Is Transforming Digital Finance: A Comparative Analysis Between Türkiye and Germany

Rahmatullah Mohammed[1]

Abstract

Blockchain technology has progressed from a novel notion in bitcoin to a widely used system in nations throughout the world, encompassing a wider range of industries, including agriculture. The decentralized nature of blockchain allows for increased transparency, cheaper transaction costs, and greater financial inclusion. This study reveals a comparative analysis of the applications of tokenized stocks, blockchain adoption, and fintech ecosystem between two countries, an emerging one, Türkiye, and a developed country, Germany, based on studies conducted between 2019 and 2025, to see how they differed in their approaches to implementation as Germany taking the lead in government adoption and Türkiye being the lead in fast adoption of innovations, as well as how each of the countries can learn from the other by presenting similarities and differences that will guide this research paper.

Keywords: Blockchain, Digitalization, Defi, Germany, Türkiye, Digital Finance, Regulation, Tokenized Stocks, Fintech.

JEL Codes: G2, G3.

Blokzincir Dijital Finansı Nasıl Dönüştürüyor: Türkiye ve Almanya Arasında Karşılaştırmalı Bir Analiz

Özet

Blokzincir teknolojisi, Bitcoin ile ortaya çıkan yeni bir kavram olmaktan çıkıp, tarım da dahil olmak üzere çok daha geniş bir sektör yelpazesini kapsayarak dünya genelinde yaygın kullanılan bir sistem haline gelmiştir. Blokzincirin merkeziyetsiz yapısı; şeffaflığın artmasına, işlem maliyetlerinin düşmesine ve finansal kapsayıcılığın güçlenmesine olanak tanır. Bu çalışma, 2019-2025 yılları arasında gerçekleştirilen araştırmalara dayanarak; gelişmekte olan bir ülke olan Türkiye ile gelişmiş bir ülke olan Almanya arasındaki tokenize hisse senedi uygulamaları, blokzincir benimsenmesi ve fintek ekosistemine yönelik karşılaştırmalı bir analiz sunmaktadır. Çalışma, Almanya'nın devlet düzeyinde benimseme, Türkiye'nin ise yeniliklerin hızlı adaptasyonu konularında öncü rollerini ele alarak, bu iki ülkenin uygulama yaklaşımlarındaki farklılıkları incelemekte; ayrıca sunduğu benzerlikler ve farklılıklar aracılığıyla ülkelerin birbirinden neler öğrenebileceğini ortaya koyarak araştırmaya yön vermektedir.

Anahtar Kelimeler: Blokzinciri, Dijitalleşme, DeFi, Almanya, Türkiye, Dijital Finans, Düzenleme, Tokenize Hisseler, Fintech.

JEL Kodları: G21, O33, F36.

1. INTRODUCTION

The potential paradigm shifts in banking, investment, and ownership brought about by blockchain (Goldestin et.al., 2024), along with the challenges related to regulation, sustainability, and scalability, has also been a center of attention for the past decade. Blockchain, therefore, represents the opportunity of democratizing finance and holds the possibility of adapting to the global finance landscape (Goghie, 2024), The digital revolution has drastically affected the way financial transactions are conducted, opening a new chapter in the decentralized impact on global markets. Blockchain, which was conceptualized in relation to cryptocurrencies such as Bitcoin, is one of the critical building blocks of the digital finance revolution (Osmani et al., 2021). Being decentralized, immutable, and transparent, it has introduced a paradigm shift in financial systems, transitioning from centralized to more efficient and inclusive models (Risman, 2021). Financial institutions, start-ups, and governments have begun to

[1] Rahmatullah Mohammed, M.Sc. Student in Finance and Banking, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
e-mail: rahmatullahmohammed2003@gmail.com, ORCID:0009-0005-9784-6866

express interest in the application of blockchain to enhance transparency, combat fraud, and maximize the efficiency of payments and exchanges. Digital finance, with such tools, is changing the concept of global inclusion, offering safe and secure avenues to the unbanked (Quader and Cek, 2024). With the digitization of the global economy, insight into the financial revolution brought by blockchain in cryptocurrency, along with tokenized stocks, is highly imperative. (Yaqoob et al., 2021). As a developed economy, Germany has experienced extensive integration of blockchain technology in their financial and technological environments. The country has a stable regulatory system, developed fintech infrastructure, and high levels of digital literacy, which is ideal for implementing blockchain technology (Hornuf and Mattusch, 2025). On the other hand, Türkiye, considered a developing economy, offers a distinct scenario with a rapidly expanding economy, a young and tech-savvy population, and a constantly shifting regulatory landscape (Kaygin et al., 2020; Canatan, 2024). These circumstances create a unique opportunity for comparison with Germany in terms of implementing blockchain in addition to tokenized assets and Fintech infrastructure. (Karaömer, 2021). With this assessment in mind, this paper responds to such central inquiries: What is the divergence in the process for implementing blockchain in emerging and developed economies? What is the implication for regulation and fintech infrastructure in this implementation? This paper applies both institutional and technological approaches to this subject matter with a comprehensive assessment of how blockchain technology is shifting and will continue to shift the construction of new financial systems.

2. METHODOLOGY

This study adopts a comparative qualitative research design that focuses on secondary data regarding regulations and adaptations of blockchain technology, tokenization of assets and the fintech ecosystem between Türkiye and Germany, the goal is to find similarities, differences, strengths and weaknesses between them in such sectors to propose a strategy that will enhance the financial ecosystem by using academic journals, legal and policy documents, case studies and institutional reports.

3. LITERATURE REVIEW

3.1. Applications in Türkiye

Türkiye is an emerging country that has a powerful development strategy and can adapt quickly to changes in foreign markets (Somuncu, 2025), furthermore, the steady growth of Türkiye in the EDGI-Government Development Index, see figure 1, and the country ranking the 22nd among the countries listed in E-participation (UN, 2024), gives a good first impression of how Türkiye is catching up with the digital competitiveness. Yet, with respect to blockchain technology, there's still an issue regarding governmental adaptation (Özturan et al., 2019), many researchers present innovative ideas regarding blockchain implementation, like E voting (Bulut et al, 2020), land registry using blockchain technology (Mendi, 2020), and agriculture supply chain movements (Gülen and Karaağaç, 2022), in addition to well-known banks like GarantiBBVA (Temuçin, 2024) and Akbank that launched pilot projects using blockchain technology, mainly using it for the purpose of transferring money and doing different financial transactions from Türkiye to other countries and as public awareness has an advanced indicator with the Central Bank of the Republic of Türkiye (CBRT) has promoted further research into blockchain usages and tokenized assets (Duğru, 2024), legal frameworks and human expertise regarding these innovations are limited and still unclear (Degerli, 2019).

3.2. Application in Germany

Germany stands as one of the European leaders in transitioning to the digital financial era, as the Electronic Security Act (eWpG, 2021) and BaFin-approved STOs (Security Token Offerings) that were enacted in Germany and promoted an enhanced and competitive market framework for the German financial position in the foreign market with a high level of survival. Furthermore, in 2024, it ranked 12th in the E-Government development index and 4th in E-Participation among the world's countries, see figure 2. Germany succeeded in having a powerful fintech infrastructure that promoted the adoption

of many different innovations in terms of blockchain technology (Nayak et al., 2020), which was integrated into the banking system, helping mainly in payments, logistics (Bartsch and Winkler, 2020), and renewable energy management in firms like InnogySE. They also used blockchain technology in Supply Chains (Toedorescu and Korchagina, 2021), as BayerAG did to improve traceability and monitor worldwide shipments. Its model represents the Superstar Firms Phenomena, which is based on manufacturing innovations rather than just a platform monopoly, being a part of a broader ecosystem supported by digital infrastructure (Brandl, 2020).

4. Comparative analysis between Germany and Türkiye

When putting both countries against each other, we can find a gap to be filled, to start, Germany and Turkiye differentiate in the blockchain adoption, fintech infrastructure, and tokenized assets making Germany a pioneer, with their efforts represented in adapting to the new digital trend and exemplifies clear regulation and implementation regarding blockchain development to enhance efficiency and transparency (Schneck, 2023), were they use the blockchain technology in more than one sector, in addition to the financial sector were they enforced DLT-Distributed Ledger Technology based platforms for securities issuance as tokenized assets in more than one financial institution like BaFin, Bundesbank and their stock market named as Deutsche Börse (Wu et al., 2024). Collectively, Germany follows a top-down approach where there's a transparent and clear governmental adaptation, even though the regulations are strict, but it has made them the champions in our analysis today (Lehmann, 2023). On the other hand, we have Türkiye, it's well-known for their agile adaptation and big public awareness about blockchain, tokenized stocks and fintech infrastructure (Yazıcı, 2019), with a an excellent number of innovative research papers that present ways in which Türkiye can adapt to such practices, highlighting how individuals are publishing the awareness among the society in different sectors like real estate, e voting, and agriculture, in addition to their blockchain technology implementation in financial institution in their big banks (Toraman, 2022), digital inclusion in Türkiye focuses on the finance sector rather than public administration and comes mainly from private firms. However, Türkiye's regulation framework and governmental inclusivity in Blockchain adaptation, tokenization of assets, and fintech ecosystem still not clear, like the case in Germany; there's also uncertainty revolving around such technologies especially when faced with limited human capital to such sectors, and with the absence of regulatory adaptation, it's hard for the investor to put trust in such products. Yet, CBRT now is promoting more to focus on R&D's regarding this sector to handle such big innovations and direct them into public benefit (Asfuroglu, 2024), this movement is a proof how Türkiye is emerging with such new concepts to be able to compete in the word financial market which begins by a few individuals who shed light on a new ecosystem, highlighting Türkiye's Bottom-up model.

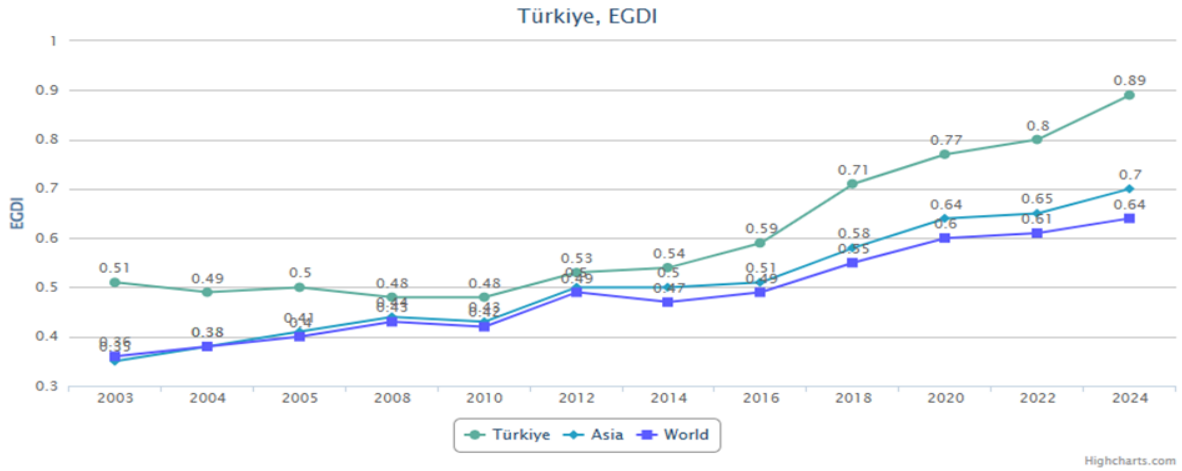


Figure 1. Türkiye's EGDI development between 2003 and 2024. Adapted from E-Government Survey 2024, United Nations

Source: https://publicadministration.un.org/egovkb/en-us/

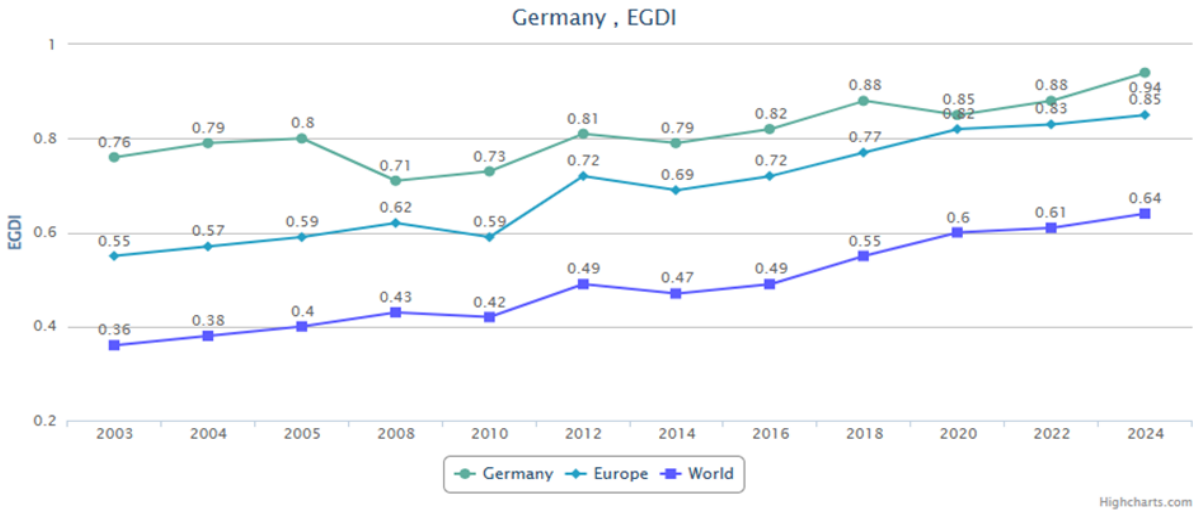


Figure 2. Germany's EGDI development between 2003 and 2024. Adapted from E-Government Survey 2024, United Nations.
Source: https://publicadministration.un.org/egovkb/en-us/

Table 1. A comparative table: Türkiye vs. Germany in Blockchain & Digital Finance

Category	**Germany**	**Türkiye**
Fintech Adoption	Stable And Mature	High Public Awareness, Growing Fast.
Tokenized Assets	Fully Regulated STOs and Tokenized Securities	Limited Experimentation
Regulations	Detailed Regulation (BaFin, eWpG)	Unclear With High Uncertainty
Government Involvement	Powerful Government Leadership and Long-Term Strategy	CBRT has recently increased R&D
Cases Applied	Banking, Logistics, Renewable Energy, Supply Chain	Banking Pilots, E-Voting Proposals, Land Registry Ideas, Agriculture
Strengths	Legal Certainty, High Trust, Advanced Digital Infrastructure	Fast Adoption and Innovation, Young-Savvy Population
Weaknesses	Slow Experimentation Due to Strict Rules	Low Legal Involvement and Limited Human Capital

5. Findings and Recommendations

Despite the differences between Türkiye's and Germany's digital infrastructure, blockchain adoption, and tokenized assets, each can learn from the other something valuable. Now for what Germany can learn from Türkiye is that Türkiye proved itself by being the most agile and much more flexible when introduced to new technologies in addition to their experimental approaches were Turkish financial institutions showed a fast movement in testing pilot projects specifically in cross-border transactions, Germany's fintech structure can learn from Türkiye's its encouragement to innovations and their quick adaption whereas it's still an emerging country.

And what Türkiye can learn from Germany is their regulation clarity and governmental involvement and motivations for presenting more studies in fintech and the tokenization of assets, not leaving it only to the private sector, which will attract more investors since it will rise the confidence among the public, in addition to involving blockchain technology in more sectors than just financial. Such diverse aspects can make Türkiye's ecosystem to take a step further in establishing a steady financial environment making the digital finance industry more accessible to foreign investment (Kömürcüoğlu, 2025).

Finally, a paradigm that can be considered a combination between legal clarity Pagin Germany and innovative adaptability in Türkiye would give a well-rounded strategy for long-term growth related to blockchain. Cross-border work, collaborations related to research, and conversations about regulations between these two countries might enhance competitiveness in this area for Europe and Eurasia on a global scale.

REFERENCES

Asfuroglu, D. (2024). Central bank digital currency in an emerging market economy: Case of the Central Bank of the Republic of Türkiye, *Current Research in Social Sciences*, 10(1), 62-74.

Bartsch, D. and H. Winkler (2020). Blockchain technology in Germany: An excerpt of real use. International Conference of Logistics (pp. 699-735). Hamburg

Brandl, B. (2020). Where did FinTechs come from, and where do they go? The transformation of the financial industry in Germany after digitalization. *Frontiers In Artificial Intelligence*, 3, 1-12.

Bulut, R., A. Kantarcı, S. Keskin and Ş. Bahtiyar (2020). Blockchain-Based Electronic Voting System for Elections in Turkey. Istanbul Technical University.

Canatan, B. (2024). Dynamics of Digital Financial Inclusion in Türkiye, *Istanbul Business Research*, 53(2), 185-200.

Değerli, K. (2019). Regulatory Challenges and Solutions for Fintech in Turkey, *Procedia* Computer Science, 158: 929-937.

Duğru, M. (2024). Preparing for the future of central bank digital money in Türkiye In Global Developments in Central Bank Digital Currency. IGI Global, 146-159.

Goghie, A. (2024). Tokenization and the banking system: Redefining Authority in the Blockchain Era, *Competition & Change*, 28(5), 663–682.

Goldestin, I., D. Gupta and R. Svewrchkov (2024). Utility Tokens as a Commitment to Competition. *The Journal of Finance*, 79(6), 4197-4233.

Groene, N. and L. Schneck (2023). Covering digital health applications in the public insurance systhow to foster innovation in patient care while mitigating financial risks—evidence from Germany. *Frontiers Digital Health*, 5:1217479

Gülen, K. G. and A. Karaağaç (2022). Agricultural Food Supply Chain with Blockchain Technology: A Review on Turkey, *Global Strategic Management*, 16(2), 013-028.

Hornuf, L. and M. Mattusch (2025). Fintech startups in Germany: firm failure, funding. *The European Journal of Finance*, 1-45.

Karaömer, Y. (2021). An overview of financial technology sector in Turkey. *Journal of Politics, Economy and Management*, 4(2), 128-139.

Kaygin, E., Y. Zengin, E. Topçuoğlu and S. Özkes (2020). The Evaluation of Block Chain Technology within the Scope of Ripple and Banking Activities. *Journal of Central Banking Theory and Practice*, 10(3), 153-167.

Kömürcüoğlu, Ö. F. (2025). The impact of FinTech on economic growth in Türkiye: A novel FinTech

index approach. *Journal of Economics and Management*, 47(1), 543-568.

Lehmann, P. P. (2023). Measuring the Success of Digital Transformation in German SMEs. *Journal of Small Business Strategy*, 33(1), 1-19.

Mendi, A.F., Ö. Demir, K.K. Sakaklı and A. Çabuk (2020). A New Approach to Land Registry System in Turkey: Blockchain-Based System Proposal. American Society for Photogrammetry and Remote Sensing, 86(11), 701–709.

Nayak, K., P. Singh and P. Dave (2020). Does Data Security and Trust Affect the Users of Fintech? *International Journal of Management*, 12(1), 191-206.

Osmani, M., R. Elhaddadeh, N. Hindi, M. Janssen and V. Weerakkody (2021). Blockchain for next generation services in banking and finance: cost, benefit, risk and opportunity analysis. *Journal of Enterprise Information Management*, 34(3), 884-899.

Özturan, M., İ. Atasu and H. Soydan (2019). Assessment of Blockchain Technology Readiness Level of Banking Industry: Case of Turkey. *International Journal of Business Marketing and Management*, 4(12), 01-13.

Quader, K.S. and K. Cek (2024). Influence of blockchain and artificial intelligence on audit quality: Evidence from Turkey, *Heliyon*, 10(9), e30166

Risman, A, B. Mulyana, B. A. Silvatika and A. S. Sulaema (2021). The effect of digital finance on financial stability. Management Science Letters, 1979-1984.

Somuncu, K. (2025). The Impact of Financial Inclusion on Economic Growth: The Case of Turkey. *Journal of Administrative Sciences*, 23(55), 310-335.

Temuçin, T. S. (2024). Role Of Machine Learning in Internal Fraud Detection. TİDE Academia Research, 6(2), 151-172.

Toedorescu, M. and E. Korchagina (2021). Applying Blockchain in the Modern Supply Chain. *Journal of Open Innovation: Technology, Market, and Complexity.* 7(1), 2-18.

Toraman, Y. (2022). Interest-Free Finance Model by Using Blockchain-Based Company Tokens: Research on Digital Turkish Lira (DTL) and Borsa Istanbul with Technology Acceptance Model (TAM). *Emerging Market Journal*, 12(2),

UN. (2024). Department of Economic and Social Affairs. Retrieved from UN E-Government Survey: https://publicadministration.un.org/egovkb/en-us/Data/Country-Information/id/176-Trkiye

Wu, H., Q. Yao, Z. Liu, B. Huang, Y. Zhuang, H. Tang and E. Liu (2024). Blockchain for finance: A survey. The Institution of Engineering and Technology, 4(2), 101-123.

Yaqoob, I., K. Salah, R. Jayaraman and Y. Al-Hammadi (2021). Blockchain for Healthcare Data Management: Opportunities, Challenges, and Future Recommendations, *Neural Computing and Application*, 34(14), 11475-11490.

Yazıcı, S. (2019). The Analysis of Fintech Ecosystem in Turkey. Journal of Business, Economics and Finance, 8(4)I 188-197.

Moment Inequality Estimators in Competition Analysis: A Practitioner's Guide with Synthetic Data Applications

Fatih Cemil Özbuğday[1]

Abstract

Moment inequality estimators have become a central tool in empirical industrial organization, yet they remain less familiar to many competition policy practitioners. This paper provides a self-contained tutorial for antitrust economists, competition authorities, and economic consultancies. It explains the shift from point identification to set identification, presents the mathematical foundations in competition-relevant terms, and develops two synthetic applications. The first is a two-player entry game with multiple equilibria, following Ciliberto and Tamer (2009), where the identified set for competitive-effect parameters is constructed from probability bounds over the multiplicity region. The second is a stylized logit demand example with an unknown outside-good share, where valid unconditional moments bound the price coefficient, and the resulting confidence set is projected into post-merger price effects. The paper also discusses extensions to collusion screening, market definition, and dynamic entry, and it explains how to communicate set-identified results in enforcement proceedings.

Keywords: Moment Inequalities, Partial Identification, Competition Policy, Entry Games, Merger Simulation.

JEL Codes: C12, C13, C51, L13, L40, L41.

Rekabet Analizinde Moment Eşitsizliği Tahmincileri: Sentetik Veri Uygulamalı Uygulamacı Kılavuzu

Özet

Beklem eşitsizliği tahmincileri, ampirik endüstriyel organizasyon alanında temel bir araç haline gelmiş olsa da rekabet politikası uygulayıcıları için büyük ölçüde erişilemez durumdadır. Bu makale; antitröst ekonomistleri, rekabet otoriteleri ve ekonomik danışmanlık firmaları için beklem eşitsizliği yöntemleri üzerine kendi içinde bütünlüğe sahip kapsamlı bir kılavuz sunmaktadır. Çalışmada, nokta özdeşlemesinden (point identification) küme özdeşlemesine (set identification) yönelik kavramsal geçiş açıklanmakta, matematiksel temeller rekabet bağlamına uygun terimlerle verilmekte ve Canay, Illanes ve Velez'den (2023) uyarlanan bir uygulama şablonu sunulmaktadır. Söz konusu yöntemler, temel antitröst senaryolarında iki sentetik veri uygulaması aracılığıyla örneklendirilmektedir. (i) İki oyunculu bir pazara giriş oyunu: Ciliberto ve Tamer (2009) izlenerek kurgulanan ve birden fazla dengenin (çoklu denge) oluşabildiği bir modeldir. Bu modelde rekabetçi etki parametreleri, çokluluk bölgesi (multiplicity region) üzerinde uygun olasılık sınırları kurularak tanımlanır. (ii) Stilize edilmiş bir logit talep örneği: Dışsal mal payının (outside-good share) bilinmediği bir modeldir. Bu örnekte, geçerli koşulsuz beklemler fiyat katsayısına bir sınır çizer; ardından elde edilen bu güven kümesi kullanılarak birleşme sonrası fiyat etkileri tahmin edilir. Her iki uygulama da geçerli kritik değer hesaplamaları ile birlikte Python'da hayata geçirilmiştir. Son olarak, yöntemlerin gizli anlaşma taraması, pazar tanımı ve dinamik pazara giriş konularındaki uzantıları tartışılmakta; beklem eşitsizliği yöntemlerinin hangi durumlarda standart tekniklerden daha iyi performans gösterdiği ve kümesel olarak tanımlanmış sonuçların hukuki yaptırım süreçlerinde nasıl aktarılabileceği ele alınmaktadır.

Anahtar Kelimeler: Moment Eşitsizlikleri, Kısmi Tanımlama, Rekabet Politikası.

JEL Kodları: C12, C13, C51, L13, L40, L41.

[1] Fatih Cemil Özbuğday, Professor of Economics, Department of Economics, Faculty of Political Sciences, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
email: fcozbugday@aybu.edu.tr, ORCID: https://orcid.org/0000-0001-5841-4343.

1. INTRODUCTION

Competition economists routinely face a fundamental tension. The structural models they use to evaluate mergers, detect collusion, or assess market power often require strong assumptions about equilibrium selection, functional forms, or data completeness. These assumptions may be difficult to defend in adversarial proceedings. When a judge asks whether a merger will raise prices by 5% or 15%, the honest answer is frequently that the data support a range of values, not a single point estimate. Yet the standard econometric toolkit is designed to deliver point estimates. Practitioners are therefore forced to make identification assumptions they cannot justify or to present false precision.

Moment inequality estimators offer an alternative. Rather than imposing strong assumptions to achieve point identification, these methods work with weaker, more defensible assumptions that identify a *set* of parameter values consistent with the data. The result is not a single estimate of the merger price effect, but a *range*, say 3% to 11%, derived from assumptions that both sides in a proceeding can accept. This shift from point estimation to set estimation may initially seem like a step backward. In fact, it is a step toward greater credibility. The reported bounds are robust to precisely those modeling choices that opposing experts would attack.

Despite their intellectual appeal and growing use in academic industrial organization (IO), moment inequality methods remain underutilized by competition practitioners. The literature includes the Handbook of Industrial Organization survey by Kline et al. (2021), the applied user's guide by Canay et al. (2024), and foundational theoretical contributions by Chernozhukov et al. (2007), Andrews and Soares (2010), and Bugni (2010). Broader surveys of partial identification are provided by Tamer (2010) and Molinari (2020), and Ho and Rosen (2017) discuss the benefits and challenges of these methods in applied work. Canay et al. (2024) provide a valuable general template, but these resources are written primarily for academic econometricians and applied IO researchers rather than for economists who evaluate mergers, screen for cartels, or assess market power in enforcement proceedings.

This paper complements those resources in three ways. First, it frames moment inequality methods in competition policy language. It translates "identified sets" into "ranges of competitive effects consistent with the data" and "confidence sets" into "bounds that hold under sampling uncertainty." Second, it develops two synthetic examples in core antitrust settings: entry games with multiple equilibria and demand estimation with unknown outside-good shares. Third, it offers guidance on when these methods add value over standard techniques and how to communicate set-identified results to non-technical decision-makers.

The paper assumes familiarity with graduate-level microeconomics and basic econometrics (OLS, IV, GMM), but not with the partial identification literature. It serves three audiences. These are (i) economists at competition authorities; (ii) economists at consulting firms; and (iii) academic researchers in IO and competition economics who wish to apply these methods in policy-relevant work.

The paper proceeds as follows. Section 2 explains the conceptual shift from point identification to partial identification, using competition-relevant examples. Section 3 presents the mathematical foundations at a level appropriate for practitioners. Section 4 provides a four-step implementation guide. Sections 5 and 6 present two synthetic data applications, namely entry games and demand estimation with merger simulation. Section 7 discusses extensions to collusion screening, market definition, and dynamic entry. Section 8 discusses computational considerations. Section 9 addresses when practitioners should use these methods and how to communicate results. Section 10 concludes.

2. FROM POINT ESTIMATION TO SET ESTIMATION: WHY MOMENT INEQUALITIES?

2.1. The Identification Problem in Competition Analysis

Consider a simple example familiar to any antitrust economist. Two firms, an incumbent and a potential entrant, simultaneously decide whether to operate in a market. Firm i's profit from entry depends on market size X, its own cost parameters θ_i, and whether the other firm enters. A standard approach would estimate the entry game by maximum likelihood, but this requires specifying which equilibrium is played when multiple equilibria exist. In many market configurations, both "firm 1 alone" and "firm 2 alone" are Nash equilibria, and the econometrician cannot determine which one generated the data without an arbitrary equilibrium selection rule.

This is not an academic curiosity. Multiple equilibria are pervasive in the models that competition economists use. Examples include entry games, pricing models with strategic complementarities, network formation models, and dynamic investment games. Any model in which firms' decisions depend on other firms' decisions, that is, any model of strategic interaction, can exhibit multiplicity. Foundational treatments of identification in such models include Berry and Tamer (2006) for oligopoly entry and Haile and Tamer (2003) for incomplete models of auctions.

The traditional response has been to impose an equilibrium selection assumption. One may assume that the Pareto-dominant equilibrium is always played or that firms coordinate on one specific equilibrium. But in adversarial proceedings, the opposing expert will challenge this assumption, and there is no empirical way to determine which selection rule is correct.

2.2. The Moment Inequality Solution

Moment inequality estimators resolve this problem by working only with implications of the model that hold across *all* equilibria. Instead of asking "which equilibrium is played?", they ask "what can we learn without answering that question?"

The answer, it turns out, is often quite a lot. While the model does not pin down a unique prediction for observed market outcomes, it does restrict the set of outcomes to a feasible region. These restrictions take the form of inequality constraints on moments of the data

$$\mathbb{E}[m(X,\theta)] \leq 0 \tag{1}$$

rather than the equality constraints familiar from GMM

$$\mathbb{E}[m(X,\theta)] = 0 \tag{2}$$

2.3. Identified Set versus Confidence Set

Two concepts must be carefully distinguished, as they serve different roles in enforcement proceedings.

The *identified set* Θ_I is the set of parameter values consistent with the population distribution of the data and the maintained model assumptions

$$\Theta_I = \{\theta \in \Theta : \mathbb{E}[m(X,\theta)] \leq 0\} \tag{3}$$

This set reflects what the model and data restrictions allow *in the population*. It captures model uncertainty (which equilibrium? which functional form?) but not sampling uncertainty.

The *confidence set* CS_n additionally accounts for sampling uncertainty. It is a random set, depending on the observed sample, that covers the identified set with probability at least $1 - \alpha$

$$\Pr(\Theta_I \subseteq CS_n) \geq 1 - \alpha \tag{4}$$

In practice, the confidence set is always weakly larger than the identified set. For policy communication, the distinction matters. The identified set answers "what does the model allow?", while the confidence set answers "what can we reliably conclude from this sample?"

2.4. Competition-Relevant Sources of Moment Inequalities

In competition analysis, moment inequalities arise from several sources.

2.4.1. Multiple Equilibria in Entry Games

When the equilibrium selection rule is unknown, the probability of observing a particular market structure is bounded above and below by the probabilities under different equilibria. The upper bound for an outcome includes the probability mass from the "multiplicity region" of the error space where that outcome is *one of* the equilibria; the lower bound excludes this mass. These bounds generate moment inequalities. Under independence of the profit shocks, the resulting identification regions are sharp in the sense of Beresteanu et al. (2011).

2.4.2. Revealed Preference in Pricing and Product Choice

If a firm chose price $\boldsymbol{p}$ over alternative $\boldsymbol{p}'$, it must be that profits at $\boldsymbol{p}$ weakly exceed profits at $\boldsymbol{p}'$. This generates the inequality $\boldsymbol{\pi}(\boldsymbol{p}, \boldsymbol{\theta}) \geq \boldsymbol{\pi}(\boldsymbol{p}', \boldsymbol{\theta})$. Since the inequality holds for all alternatives $\boldsymbol{p}'$, a large system of moment inequalities emerges (Pakes, 2010; Pakes et al., 2015).

2.4.3. Incomplete Data

When market shares are observed with error, when the "outside good" share is unknown (Bugni et al., 2026), or when data are available only in aggregated form, equality-based moment conditions become inequality-based. This outside-share problem is especially relevant for merger simulation because demand estimates can be sensitive to ad hoc market-size assumptions.

2.4.4. Strategic Interaction under Incomplete Information

When firms have private information about costs or demand, and multiple Bayesian Nash equilibria exist, the observable implications are characterized by moment inequalities (Xiao, 2018).

2.5. The Practitioner's Value Proposition

For the competition practitioner, moment inequalities offer three concrete advantages.

First, *robustness in adversarial settings*. Bounds derived from moment inequalities survive challenges to equilibrium selection, functional form, and data completeness. When the opposing expert argues that a different equilibrium selection rule would yield different results, the practitioner can respond that their bounds hold regardless of the selection rule.

Second, *transparency about uncertainty*. Rather than presenting a point estimate with a confidence interval that reflects only sampling uncertainty (not model uncertainty), moment inequalities produce

confidence sets that reflect both sources of uncertainty. The width of the confidence set communicates how much the data can tell us, given the maintained assumptions.

Third, *disciplined weakening of assumptions*. Practitioners can start with strong assumptions yielding point estimates, then systematically relax assumptions and observe how the identified set widens. This "robustness exercise" reveals which assumptions drive the results. That information is valuable for both the practitioner and the decision-maker.

3. MATHEMATICAL FOUNDATIONS FOR COMPETITION PRACTITIONERS

3.1. The General Framework

We observe data $\{X_i\}_{i=1}^n$ drawn from a distribution F_0, and we have a vector-valued function $m(X,\theta)$ mapping data and parameters to $\mathbb{R}^J$, where J is the number of moment inequalities. The identified set is

$$\Theta_I = \{\theta \in \Theta : \mathbb{E}_{F_0}[m_j(X,\theta)] \leq 0 \text{ for all } j = 1, \dots, J\} \tag{5}$$

In practice, we replace the population expectation with the sample analog

$$\bar{m}_j(\theta) = \frac{1}{n}\sum_{i=1}^n m_j(X_i,\theta) \tag{6}$$

The statistical problem is to construct a confidence set CS_n that covers the identified set Θ_I with probability at least $1-\alpha$. This is achieved by *test inversion*. For each candidate parameter value θ, test the null hypothesis $H_0: \theta \in \Theta_I$ against $H_1: \theta \notin \Theta_I$, and include θ in the confidence set if the test does not reject.

3.2. Test Statistics

The modified method of moments (MMM) statistic is

$$T_n(\theta) = \sum_{j=1}^J \left[\frac{\sqrt{n}\,\bar{m}_j(\theta)}{\hat{\sigma}_j(\theta)}\right]_+^2 \tag{7}$$

where $[x]_+ = \max(x,0)$ and $\hat{\sigma}_j(\theta)$ is the sample standard deviation of $m_j(X_i,\theta)$. This statistic measures how much the sample moments violate the inequality constraints. When all sample moments satisfy $\bar{m}_j(\theta) \leq 0$, the statistic is zero and θ is not rejected.

The quasi-likelihood ratio (QLR) statistic is

$$QLR_n(\theta) = \inf_{t \leq 0}\left[\sqrt{n}(\bar{m}(\theta) - t)\right]' \hat{\Sigma}^{-1}(\theta)\left[\sqrt{n}(\bar{m}(\theta) - t)\right] \tag{8}$$

where $\hat{\Sigma}(\theta)$ is the variance-covariance matrix of the moment functions. The QLR statistic projects the standardized sample moments onto the nearest point in the non-positive orthant and measures the squared distance.

3.3. Critical Values

The critical value $c_n(\theta, 1-\alpha)$ must ensure uniform size control

$$\sup_{\theta \in \Theta_I} \Pr\left(T_n(\theta) > c_n(\theta, 1-\alpha)\right) \leq \alpha \tag{9}$$

Three approaches are standard.

Least favorable (LF) critical values assume all moment inequalities bind simultaneously, i.e., $\mathbb{E}[\boldsymbol{m_j}(\boldsymbol{X},\boldsymbol{\theta})] = \mathbf{0}$ for all $\boldsymbol{j}$. Under this configuration, the MMM statistic converges to $\boldsymbol{T}^* = \sum_{\boldsymbol{j}} [\boldsymbol{Z}_{\boldsymbol{j}}^*]_+^{\mathbf{2}}$ where $\boldsymbol{Z}^* \sim \boldsymbol{N}\left(\mathbf{0}, \widehat{\boldsymbol{\Omega}}(\boldsymbol{\theta})\right)$ and $\widehat{\boldsymbol{\Omega}}(\boldsymbol{\theta})$ is the estimated *correlation* matrix of the standardized moments. The LF critical value is the $\mathbf{1} - \boldsymbol{\alpha}$ quantile of $\boldsymbol{T}^*$. It is computed by the following simulation.

In practice, the LF critical value is obtained by drawing $\boldsymbol{B}$ realizations from $\boldsymbol{N}(\mathbf{0}, \widehat{\boldsymbol{\Omega}})$, computing the simulated MMM statistic for each draw, and taking the $(\mathbf{1} - \boldsymbol{\alpha})$ quantile of the simulated statistics.

This is *not* a χ_J^2 quantile unless the moments are uncorrelated. The LF approach is conservative because the confidence set may be larger than necessary, but it guarantees correct coverage.

Generalized moment selection (GMS) critical values (Andrews and Soares, 2010) use the data to classify each moment as "near binding" or "far from binding." Moments with $\sqrt{\boldsymbol{n}}\bar{\boldsymbol{m}}_{\boldsymbol{j}}(\boldsymbol{\theta})/\hat{\boldsymbol{\sigma}}_{\boldsymbol{j}}(\boldsymbol{\theta}) \ll \mathbf{0}$ (far below zero) are clearly slack and contribute no information; they are effectively set to $-\infty$ in the simulation. Moments near zero are treated as potentially binding. This yields tighter confidence sets and maintains correct asymptotic coverage. Specifically, the GMS approach replaces the zero mean in the LF simulation with a data-dependent shift

$$\varphi_j = \begin{cases} 0 & \text{if } \sqrt{n}\bar{m}_j(\theta)/\hat{\sigma}_j(\theta) \geq -\kappa_n \\ -\infty & \text{if } \sqrt{n}\bar{m}_j(\theta)/\hat{\sigma}_j(\theta) < -\kappa_n \end{cases} \tag{10}$$

where κ_n diverges slowly (e.g., $\kappa_n = \sqrt{\ln n}$). The theory requires $\kappa_n \to \infty$ and $\kappa_n/\mathrm{n} \to 0$; $\sqrt{\ln n}$ is the standard default because it satisfies both conditions and tends to work well in practice. The GMS critical value is then the $(1 - \alpha)$ quantile of the simulated statistic.

Bootstrap critical values (Bugni, 2010) use resampling to approximate the distribution of the test statistic under the null. This provides an alternative that can adapt to the local structure of the moment functions.

3.4. From Entry Games to Moment Inequalities: A Worked Derivation

Consider a market with two potential entrants, $i \in \{1,2\}$. Each firm decides simultaneously whether to enter ($Y_i = 1$) or stay out ($Y_i = 0$). Firm i's latent payoff from entry in market ℓ is

$$\pi_{i\ell} = \alpha_i + \beta_i X_\ell + \delta_i Y_{-i,\ell} + \varepsilon_{i\ell} \tag{11}$$

where X_ℓ is market size, $\delta_i < 0$ captures the competitive effect of the rival's entry, and $\varepsilon_{i\ell} \sim N(0,1)$ independently. Define four entry thresholds

$$t_1^0 = -(\alpha_1 + \beta_1 X) \qquad t_1^1 = -(\alpha_1 + \beta_1 X + \delta_1) \tag{12}$$

$$t_2^0 = -(\alpha_2 + \beta_2 X) \qquad t_2^1 = -(\alpha_2 + \beta_2 X + \delta_2) \tag{13}$$

Since $\delta_i < 0$, we have $t_i^1 > t_i^0$. It is harder to enter when the rival enters. The $(\varepsilon_1, \varepsilon_2)$ space is partitioned by these thresholds into a 3×3 grid of regions (below/between/above each pair of thresholds). Analysis of best responses in each region yields the following cases.

- **Unique equilibrium (0,0):** $\varepsilon_1 < t_1^0$ and $\varepsilon_2 < t_2^0$
- **Unique equilibrium (1,1):** $\varepsilon_1 \geq t_1^1$ and $\varepsilon_2 \geq t_2^1$

- **Unique equilibrium (1,0):** ($\varepsilon_1 \geq t_1^0$ and $\varepsilon_2 < t_2^0$) or ($\varepsilon_1 \geq t_1^1$ and $t_2^0 \leq \varepsilon_2 < t_2^1$)
- **Unique equilibrium (0,1):** ($\varepsilon_2 \geq t_2^0$ and $\varepsilon_1 < t_1^0$) or ($\varepsilon_2 \geq t_2^1$ and $t_1^0 \leq \varepsilon_1 < t_1^1$)
- **Multiple equilibria {(1,0), (0,1)}:** $t_1^0 \leq \varepsilon_1 < t_1^1$ and $t_2^0 \leq \varepsilon_2 < t_2^1$

The multiplicity region is the rectangle $[t_1^0, t_1^1) \times [t_2^0, t_2^1)$, where *both* (1,0) and (0,1) are Nash equilibria but the econometrician cannot determine which is played. The probability mass in this region, $P_M(X,\theta) = [\Phi(t_1^1) - \Phi(t_1^0)][\Phi(t_2^1) - \Phi(t_2^0)]$, is the source of partial identification.

For each outcome $(\boldsymbol{y_1}, \boldsymbol{y_2})$, define the following objects.

- $P_L(y_1, y_2 \mid X, \theta)$ denotes the probability that (y_1, y_2) is the *unique* equilibrium
- $P_U(y_1, y_2 \mid X, \theta)$ denotes the probability that (y_1, y_2) is *an* equilibrium (unique or one of several)

For outcomes not involved in multiplicity, $P_L = P_U$

$$P_L(0,0) = P_U(0,0) = \Phi(t_1^0)\Phi(t_2^0) \tag{14}$$

$$P_L(1,1) = P_U(1,1) = [1 - \Phi(t_1^1)][1 - \Phi(t_2^1)] \tag{15}$$

For outcomes involved in multiplicity, the bounds are

$$P_L(1,0) = [1 - \Phi(t_1^0)]\Phi(t_2^0) + [1 - \Phi(t_1^1)][\Phi(t_2^1) - \Phi(t_2^0)] \tag{16}$$

$$P_U(1,0) = P_L(1,0) + P_M(X,\theta) \tag{17}$$

$$P_L(0,1) = \Phi(t_1^0)[1 - \Phi(t_2^0)] + [\Phi(t_1^1) - \Phi(t_1^0)][1 - \Phi(t_2^1)] \tag{18}$$

$$P_U(0,1) = P_L(0,1) + P_M(X,\theta) \tag{19}$$

Note that the sum of the two upper bounds exceeds one whenever the multiplicity region has positive mass. This is not incoherence; it reflects incompleteness in Tamer's (2003) sense, because the model does not uniquely assign the multiplicity mass across the two single-firm outcomes. The closed-form product expressions above also rely on the independence of the two profit shocks. With correlated market-level unobservables, practitioners typically simulate the relevant probabilities from a multivariate distribution inside the estimation loop rather than factorizing them into univariate normal CDFs. The moment inequalities are

$$\mathbb{E}[\mathbf{1}\{Y = (y_1, y_2)\} - P_U(y_1, y_2 \mid X, \theta)] \leq 0 \tag{20}$$

$$\mathbb{E}[P_L(y_1, y_2 \mid X, \theta) - \mathbf{1}\{Y = (y_1, y_2)\}] \leq 0 \tag{21}$$

With four outcomes, this yields eight moment inequalities. The identified set Θ_I is the set of $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2)$ for which all eight inequalities hold in population.

4. A FOUR-STEP IMPLEMENTATION GUIDE

The implementation of moment inequality estimators in competition settings follows and adapts Canay et al. (2024). It proceeds in four steps.

Step 1. Specify the behavioral model. Define the economic model of firm behavior, such as entry, pricing, bidding, or investment. The model must imply restrictions that take the form of inequalities. In competition settings, these typically arise from profit maximization (revealed preference), equilibrium conditions with multiplicity, or data limitations.

Step 2. Derive the moment inequalities. Translate the behavioral model into formal moment inequality conditions $\mathbb{E}[m(X,\theta)] \leq 0$. This requires identifying which observable implications hold as inequalities rather than equalities. Section 3.4 illustrates this for entry games; the key is identifying the multiplicity region and computing probability bounds.

Step 3. Choose test statistic and critical value method. Select the MMM or QLR test statistic and the critical value computation method (LF, GMS, or bootstrap). For competition applications with a moderate number of moments ($J \leq 20$), GMS offers a good balance of power and reliability. The LF approach is appropriate when conservative bounds are preferred (e.g., in preliminary screening) but may produce uninformatively wide confidence sets. The critical value must account for the correlation structure of the moments. A plain χ^2_J quantile is generally incorrect.

Step 4. Compute the confidence set via test inversion. Construct a grid over the parameter space and compute the test statistic at each grid point. Compare each statistic to the critical value and retain the grid points where the test does not reject. The confidence set is the collection of non-rejected parameter values.

5. APPLICATION 1: ENTRY GAME WITH MULTIPLE EQUILIBRIA

5.1. Model and Data Generating Process

The first application uses the two-player entry game derived in Section 3.4 under the standard case of negative competitive effects, so that each firm's profit falls when its rival enters. The synthetic design contains 500 markets, a scalar market-size shifter, and independent profit shocks. The independence assumption is substantive because it allows the probability bounds to factorize into univariate normal terms. With correlated shocks, the relevant probabilities should be simulated from a multivariate distribution inside the estimation loop. When multiple equilibria exist, the data-generating process selects among them, but the econometrician does not observe or impose the selection rule.[2]

5.2. Estimation and Inference Procedure

For each candidate value of the competitive-effect parameters, the procedure computes lower and upper probabilities for the four possible market structures: no entry, both firms entering, firm 1 alone, and firm 2 alone. The lower probability uses only regions in which the outcome is the unique equilibrium. The upper probability additionally includes the multiplicity region when the outcome is one of the two single-firm equilibria. These lower and upper probabilities generate eight moment inequalities.

Inference is conducted by test inversion. The modified method-of-moments statistic measures the positive part of standardized sample-moment violations. Candidate parameter values are retained when the statistic is below a valid critical value. The least-favorable method treats all moments as potentially binding and is conservative. The generalized moment selection method uses the data to drop moments that are clearly slack, usually producing a smaller confidence set while maintaining asymptotic coverage.

[2] The Python implementation underlying both synthetic applications, including the entry-game probability bounds, the demand-side moment construction, and the merger counterfactual, is available from the author on request.

To keep the synthetic illustration transparent, the nuisance parameters governing stand-alone profitability and market-size effects are fixed at their data-generating values, and the confidence set is constructed over the two competitive-effect parameters. In empirical work, this shortcut should be replaced by a joint confidence set over the full parameter vector and projection onto the parameters or counterfactual objects of interest.

5.3. Results

With the baseline calibration, 20 of the 500 markets, or 4.0%, have multiple equilibria. The observed market-structure frequencies are 34.6% for no entry, 41.4% for firm 2 alone, 18.4% for firm 1 alone, and 5.6% for both firms entering. The modest multiplicity rate implies that most observations are governed by a unique equilibrium, but the multiplicity region remains large enough to make equilibrium selection relevant.

The least-favorable confidence set contains 87 grid points. It implies δ_1 in [-1.30, -0.40] and δ_2 in [-1.20, 0.00]. The GMS confidence set contains 75 grid points and is tighter: δ_1 in [-1.20, -0.40] and δ_2 in [-1.20, -0.10]. Both sets contain the data-generating values, $\delta_1 = -0.80$ and $\delta_2 = -0.60$. The GMS set excludes zero competitive effect for both firms at this grid resolution, whereas the least-favorable set remains more cautious for δ_2.

These results illustrate the practical value of separating model uncertainty from sampling uncertainty. The multiplicity region is small but not irrelevant. Relaxing the equilibrium-selection assumption widens the set of admissible competitive effects, yet the GMS procedure still rules out economically negligible rivalry in this synthetic example.

5.4. Competition Policy Interpretation

In this application, the competitive-effect parameter measures how much a firm's entry profit falls when the rival also enters. The estimated ranges imply moderate to strong competitive pressure. For a market of average size, the rival's presence reduces firm 1's entry probability by roughly 15 to 42 percentage points across the retained parameter values.

For an antitrust investigation, the key message is not the exact value of a structural parameter but the range of competitive discipline consistent with the data. A competition authority reviewing a merger that removes firm 2 as a potential entrant could state that, under equilibrium selection rules consistent with the observed market structures, firm 2's entry exerts meaningful discipline on firm 1. This conclusion does not require choosing among equilibria in markets where multiple entry configurations are feasible.

The multiplicity rate should be reported alongside the confidence set. A low rate indicates that partial identification only modestly widens the result relative to a point-identified model. A high rate would signal that equilibrium-selection uncertainty is central and that additional data, richer covariates, or stronger assumptions may be needed to obtain policy-relevant bounds.

6. APPLICATION 2: DEMAND ESTIMATION WITHOUT OUTSIDE GOOD SHARES

6.1. Motivation

Most structural demand models used in merger simulation require an outside-good share, that is, the fraction of potential consumers who do not purchase any product in the market. In practice, this share is often obtained from an assumed potential market size. Recent work on demand estimation without outside-good shares emphasizes how influential that assumption can be for elasticities, markups, and merger effects.

Moment inequalities can bound demand parameters without assuming that the outside share is known. The outside share is treated as a nuisance object constrained to lie in a plausible interval. To keep the exposition transparent, the application uses a simple logit demand system, profiles over the non-price parameter in the synthetic exercise, and approximates the conditional moment-inequality problem with a finite set of nonnegative basis functions of the excluded cost shifter. These are pedagogical simplifications rather than default empirical recommendations.

In practice, economists often begin with bounds on potential market size rather than with a direct bound on the outside share. Once observed inside sales are fixed, lower and upper bounds on potential market size map mechanically into lower and upper bounds on the outside share in each market. The common interval used below is therefore a synthetic shorthand for the more common exercise of translating market-size assumptions into market-specific outside-share bounds.

6.2. Model

Consider a logit demand model with J products observed across T markets. Consumer i's indirect utility from product j in market t is

$$u_{ijt} = \alpha p_{jt} + x_{jt}'\beta + \xi_{jt} + \varepsilon_{ijt} \tag{22}$$

where ε_{ijt} is i.i.d. Type-I extreme value. Given a potential market size M_t (unknown), the outside share is $s_{0t} = 1 - \sum_j s_{jt}$, where s_{jt} is product j's share of M_t. The standard BLP inversion requires observing s_{0t} to recover $\delta_{jt} = \ln(s_{jt}) - \ln(s_{0t})$.

When $\boldsymbol{M_t}$ is unknown but bounded, $\boldsymbol{s_{0t} \in [\underline{s}_0, \overline{s}_0]}$, the mean utility $\boldsymbol{\delta_{jt}}$ is bounded

$$\delta_{jt} \in \left[\ln(s_{jt}^w) - \ln(\overline{s}_0),\ \ln(s_{jt}^w) - \ln(\underline{s}_0)\right] \tag{23}$$

where s_{jt}^w is the within-purchase share. Let ξ_{jt} denote the unobserved demand shock and suppose $\mathbb{E}[\xi_{jt} \mid w_{jt}] = 0$. The implied conditional moment inequalities are

$$\mathbb{E}[\delta_{hi,jt} - \alpha p_{jt} - \beta x_j \mid w_{jt}] \geq 0 \tag{24}$$

$$\mathbb{E}[\alpha p_{jt} + \beta x_j - \delta_{lo,jt} \mid w_{jt}] \geq 0 \tag{25}$$

For inference, the conditional inequalities are converted into unconditional inequalities by multiplying them by nonnegative indicator basis functions of the cost shifter. This preserves the sign of the inequalities and avoids the invalid step of multiplying by sign-changing instruments. Conceptually, the exercise is a finite-basis approximation to the framework of Andrews and Shi (2013), who study inference for conditional moment inequalities using rich classes of nonnegative instruments. In applied work, the partition of the cost shifter should be justified and checked for robustness.

6.3. Synthetic Design and Inference Procedure

The synthetic demand application has four products in 80 markets. Product characteristics are fixed, while excluded cost shifters and demand shocks vary across markets. The econometrician observes prices and within-purchase shares, but not the outside-good share. The outside share is assumed to lie between 0.25 and 0.65, which brackets the true simulated range of approximately 0.343 to 0.619.

For each candidate price coefficient, the procedure constructs moment inequalities implied by the bounds on mean utility. A constant and three indicator bins of the cost shifter are used as nonnegative basis functions, yielding a finite set of unconditional inequalities. The modified method-of-moments

statistic and least-favorable critical values are then used to invert the test and obtain a confidence set for the price coefficient.

The merger counterfactual considers a merger between Products 1 and 2. For each non-rejected price coefficient and each admissible outside-share value on a finite grid, pre-merger utilities and marginal costs are recovered, post-merger Bertrand first-order conditions are solved, and the resulting proportional price changes are averaged across markets. The reported bounds therefore combine uncertainty about price sensitivity with uncertainty about the outside share.

6.4. Results and Merger Counterfactual

The confidence set for the price coefficient is [-1.88, -1.00], and it contains the data-generating value of -1.50. The width of the set reflects both sampling variation and the fact that the outside share is not observed. More negative values correspond to greater price sensitivity and smaller merger effects; less negative values imply more captive consumers and larger effects.

For the merger between Products 1 and 2, the projected post-merger price-effect bounds are 2.1% to 9.8% for Product 1 and 3.1% to 14.8% for Product 2. The non-merging products are only mildly affected: Product 3 rises by 0.1% to 1.5%, and Product 4 rises by 0.0% to 0.6%. Under logit demand, these smaller spillovers arise through cross-elasticities after the merging firms raise prices.

The source of width can be decomposed by holding one uncertainty fixed and varying the other. Narrower industry evidence on market size or the fraction of consumers who would not purchase any product would directly tighten the outside-share interval and, in turn, the merger price-effect bounds.

6.5. Competition Policy Interpretation

Competition authorities often use a 5% price increase as a benchmark, especially in SSNIP-style reasoning. The bounds in this application straddle that benchmark. The analysis therefore does not support a claim of robust harm, because the lower bounds are below 5%. Nor does it support robust innocence, because the upper bounds are well above 5%. The correct interpretation is that the merger may or may not cross the benchmark depending on market-size and outside-substitution assumptions.

An expert report could state that, at the 95% confidence level, the demand evidence implies post-merger price increases of 2.1% to 9.8% for Product 1 and 3.1% to 14.8% for Product 2 under outside-share values between 0.25 and 0.65. A point-identified logit model with a single assumed outside share would deliver a sharper number, but the moment inequality analysis reveals how sensitive that number is to market-size assumptions.

When the bounds straddle the legal or policy threshold, the analysis identifies exactly what additional evidence would matter. Consumer surveys, industry estimates of potential market size, or richer purchase/non-purchase data could narrow the outside-share interval. Alternatively, a richer demand model may tighten the identified set, but only by introducing stronger assumptions about substitution patterns.

7. EXTENSIONS

This section briefly describes three further applications where moment inequality methods can contribute to competition analysis. These are presented as research directions rather than fully implemented applications, as each requires careful attention to the specific identification and inference challenges involved.

7.1. Collusion Screening via Conduct Parameter Bounds

Under competitive conduct, firms' pricing decisions satisfy Cournot or Bertrand first-order conditions. Under collusion, firms internalize effects on co-conspirators' profits. The "conduct parameter" $\boldsymbol{\lambda} \in [\mathbf{0}, \mathbf{1}]$ (Bresnahan, 1982) captures the degree of internalization, with $\boldsymbol{\lambda} = \mathbf{0}$ for competition and $\boldsymbol{\lambda} = \mathbf{1}$ for perfect cartel.

In practice, $\boldsymbol{\lambda}$ may be only partially identified because marginal costs are not observed and must be backed out from the model. If cost data come with measurement error or if cost functions are misspecified, the implied $\boldsymbol{\lambda}$ lies in a set rather than at a point. Moment inequalities derived from revealed preference, namely that firms chose observed prices over deviations, can bound $\boldsymbol{\lambda}$ under weaker assumptions than full cost specification.

A proper implementation would require (i) a demand system estimated separately or jointly; (ii) instruments for prices that are excluded from the supply side; (iii) moment inequalities derived from the full profit function including unknown marginal costs; and (iv) test inversion with valid critical values. The key challenge is that the identified set for $\boldsymbol{\lambda}$ can collapse to a point if enough structure is imposed (which turns the exercise into standard conduct parameter estimation), or it can be uninformatively wide if too little structure is assumed.

7.2. Market Definition with Incomplete Substitution Data

Market definition via the SSNIP test requires estimating own-price and cross-price elasticities. When substitution to products outside the candidate market is not observed, the diversion ratios are only partially identified. Moment inequalities can bound these diversion ratios under the assumption that outside substitution lies in a specified interval, similar to the outside share approach in Application 2.

A proper implementation would extend Application 2 in three steps. First, define the SSNIP test as a function of the partially identified demand parameters. Second, compute the SSNIP outcome at each point in the identified set. Third, report whether the SSNIP conclusion is robust across the identified set. This is a sensitivity exercise, not a moment inequality estimator per se, but it can be grounded in the formal confidence set from Application 2.

7.3. Dynamic Entry with Set-Identified Barriers

In monopolization cases and merger reviews, competition authorities assess barriers to entry. Dynamic models of entry and exit (Ericson and Pakes, 1995) can identify sunk entry costs from observed entry patterns, but when the dynamic game has multiple equilibria, these costs are only set-identified. The partial identification framework of Abito and Chen (2023) uses incentive-compatibility inequalities to bound continuation payoffs and identify collusion-sustaining conditions in dynamic settings.

A proper implementation requires (i) a dynamic game with endogenous state transitions; (ii) moment inequalities derived from Bellman optimality conditions that hold across equilibria; and (iii) treatment of the computational challenge of evaluating continuation values at each grid point. This is substantially more complex than the static entry game in Application 1, and we defer it to future work.

8. COMPUTATIONAL CONSIDERATIONS

8.1. The Curse of Dimensionality

The primary computational challenge is constructing the confidence set via grid search. With $\boldsymbol{K}$ parameters of interest and $\boldsymbol{G}$ grid points per dimension, the total number of test statistic evaluations is $\boldsymbol{G^K}$. Each evaluation requires computing $\boldsymbol{J}$ moment functions across $\boldsymbol{n}$ observations and then simulating

critical values. For a two-dimensional grid with $\boldsymbol{G = 50}$, this means 2,500 evaluations, which is manageable on a laptop. For a five-dimensional parameter space, $\boldsymbol{50^5 \approx 3 \times 10^8}$ evaluations, which is infeasible without acceleration.

8.2. Practical Strategies

8.2.1. Profiling

The synthetic applications fix some nuisance parameters at their data-generating values to keep the exposition low-dimensional. That is an oracle simplification, not an empirical recommendation. In applied work, the defensible route is to construct a joint confidence set over the parameters of interest and nuisance parameters and then project onto the object one wishes to report. Canay et al. (2024, Section 7) discuss the logic and limits of profiling in moment-inequality settings.

Profiling is not automatically valid in moment-inequality models. Uniformly valid subvector inference generally requires either a joint search over the full parameter vector, additional structure that justifies profiling, or a dedicated projection-based procedure such as Bugni et al. (2017) or Kaido et al. (2019). A practitioner should therefore use joint search plus projection, or another valid subvector method, rather than plugging in a first-stage point estimate and ignoring its uncertainty.

8.2.2. Adaptive Grid Refinement

Start with a coarse grid, identify the boundary of the confidence set (where the test statistic is near the critical value), and refine the grid in that boundary region.

8.2.3. Moment Design

Application 2 uses four nonnegative basis functions for each product, generating 32 inequalities from 80 markets. That ratio is acceptable for a transparent tutorial, but in smaller samples or higher-dimensional applications a coarser basis, such as a constant plus above/below-median indicators, can improve finite-sample stability. Andrews and Shi (2013) provide the broader framework for turning conditional inequalities into many unconditional moments; in practice, one should justify the basis choice and check robustness to alternative partitions or richer sieves. If the estimated correlation matrix becomes ill-conditioned because the number of moments is large relative to the sample, the better fix is usually to simplify the moment design or enlarge the sample rather than to rely heavily on numerical repair.

8.2.4. Tuning Parameters

The GMS procedure uses $\kappa_n = \sqrt{\ln n}$ as a standard Andrews-Soares default because it satisfies $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$. Practitioners should treat this as a sensible default rather than a revealed truth and check that substantive conclusions are not driven by small changes in the tuning rule.

8.2.5. Constrained Optimization

Dong et al. (2021) reformulate the grid search as a constrained optimization problem, which can be orders of magnitude faster. Rather than evaluating the test statistic everywhere, they solve for the boundary of the confidence set directly.

8.2.6. Parallelization

Evaluations at different grid points are independent, so test inversion is naturally parallel. For scanner-sized datasets or high-dimensional grids, practitioners should combine parallel computing with adaptive grid refinement, compiled routines, or optimization-based boundary searches.

8.3. Software

A useful open-source resource is the Canay-Illanes-Velez repository, which provides implementations of the four-step procedure. More generally, GMM, optimization, and simulation tools in standard statistical software can serve as building blocks for custom moment-inequality routines.

Dedicated tooling for moment-inequality inference remains sparse relative to mainstream GMM packages. User-friendly modules for entry games, demand bounds, and generic test inversion would be valuable for competition-policy applications.

9. WHEN SHOULD PRACTITIONERS USE MOMENT INEQUALITIES?

9.1. Decision Framework

Moment inequality methods are most valuable when three conditions hold simultaneously. First, the standard point-identified model requires assumptions that are controversial or difficult to defend. Second, the moment inequality approach yields bounds narrow enough to be informative for the policy question. Third, the computational cost is manageable given the project timeline.

9.2. High-Value Applications

9.2.1. Merger Review with Contested Model Assumptions

When opposing experts disagree about functional form, equilibrium selection, or market definition, moment inequalities provide bounds that are robust to these disagreements. The merger simulation bounds from Application 2 directly serve this purpose.

9.2.2. Entry Analysis with Strategic Interaction

The entry game framework from Application 1 is directly applicable to assessing competitive effects in markets with few incumbents and potential entrants. The Ciliberto-Tamer approach has been applied to airlines, hospitals, and broadband markets. Related applications of moment-inequality and partial-identification methods include Ciliberto et al. (2021) on airline market structure, Ho and Pakes (2014) on hospital referrals, Crawford and Yurukoglu (2012) on television bundling, Eizenberg (2014) on product variety, Holmes (2011) on retail expansion, and Wollmann (2018) on commercial vehicles.

9.2.3. Preliminary Assessment with Limited Data

In the early stages of an investigation, when data are incomplete, moment inequalities can provide *some* information under weak assumptions rather than requiring the practitioner to wait for complete data or make strong assumptions.

9.3. When Standard Methods Suffice

Moment inequalities are unnecessary when the point-identified model is credible and its assumptions are uncontroversial. If both sides in a proceeding accept the BLP demand model with a known outside

share, there is no reason to replace the point estimate with wider bounds. Moment inequalities are a tool for settings where the standard assumptions are the source of disagreement.

9.4. Communicating Set-Identified Results

Presenting a range of estimates rather than a single number requires adjustment in how results are communicated. The following strategies have proven effective in practice.

Focus on the policy question, not the parameter. Decision-makers care about whether the merger will cause harm, not about the numerical value of the price coefficient. Rather than reporting "the confidence set for $\boldsymbol{\alpha}$ is $[\mathbf{-1.88}, \mathbf{-1.00}]$," report the implication. "Under all market-size assumptions consistent with industry data, the merger raises the price of the target's flagship product by between 3% and 15%. Whether this exceeds the 5% threshold depends on assumptions about the fraction of consumers who would switch to products outside the market."

This framing does three things. It answers the question the commissioner is actually asking. It identifies the key empirical uncertainty (outside substitution). It also points to the type of evidence that could resolve the uncertainty (consumer survey data on switching behavior).

Present a robustness hierarchy. Structure the findings as a progression from strong-assumption precision to weak-assumption robustness. The following progression illustrates the idea.

(a) *Standard model (strong assumptions).* "Assuming a potential market size of 1,000,000 consumers (the industry association's estimate), the predicted merger price increase is 5.5%."
(b) *Relaxed market size (moderate assumptions).* "Allowing the potential market size to range between 800,000 and 1,500,000 consumers, the predicted price increase ranges from 3.8% to 7.4%."
(c) *Unknown market size (weak assumptions).* "Using moment inequality methods that require no specific market-size assumption, only that the outside share lies between 25% and 65%, the predicted price increase ranges from 2.1% to 9.8%."

This hierarchy lets the decision-maker see exactly which assumption drives the width. If they find the market association's estimate credible, they can rely on the first statement. If they are uncertain about the market size, they can use the second or the third. They can then make their decision based on whether the range as a whole indicates harm.

Use tables that map assumptions to conclusions. A simple table can be more persuasive than a paragraph. Table 1 reports the mapping for the merger simulation in Application 2.

Table 1. Outside-share assumptions, confidence sets, and merger price effects in Application 2.

Outside share assumption	Confidence set for α	Product 1 price effect
Known at 0.45 (standard BLP)	Point estimate: -1.50	5.5%
Bounded: [0.35, 0.50]	$[-1.72, -1.25]$	4.2% to 7.1%
Bounded: [0.25, 0.65]	$[-1.88, -1.00]$	2.1% to 9.8%

Each row uses progressively weaker assumptions and produces progressively wider bounds. The decision-maker can choose the row that matches their comfort level with the identifying assumptions.

Address the "but which number do I use?" objection directly. Judges and commissioners accustomed to point estimates may resist ranges. Anticipate this objection with the following explanation. "A point estimate hides the disagreement between the two experts behind different assumptions about market

size. The range makes that disagreement visible and allows you to assess whether the disagreement matters for your decision. If the entire range implies a price increase above 5%, it does not matter which market-size assumption is correct. The conclusion is the same. If the range straddles 5%, the disagreement matters, and additional evidence on market size could resolve it."

Report what the bounds rule out, not just what they include. The confidence set excludes parameter values as well as including them. Negative findings can be as valuable as positive ones. For example, "The data rule out the possibility that consumers are highly price-insensitive ($\boldsymbol{\alpha > -1.00}$), which in turn rules out post-merger price increases above 15%. The data also rule out very high price sensitivity ($\boldsymbol{\alpha < -1.88}$), which rules out the respondent's claim that the merger would have negligible effects." Framing the exclusions as tests of the opposing party's claims can be particularly effective in adversarial proceedings.

Distinguish between what the econometrics delivers and what the policy decision requires. The econometric analysis produces a confidence set and a set of counterfactual predictions. The enforcement decision requires a judgment about whether the predicted effects constitute "substantial lessening of competition" or an equivalent legal threshold. These are separate steps. The practitioner's job is to provide the first; the decision-maker's job is to apply the second. Set identification clarifies this division of labor. It tells the decision-maker the range of effects consistent with the evidence. It leaves the judgment about whether that range crosses the legal threshold where it belongs, with the decision-maker.

10. CONCLUSION

This paper has shown that moment inequality estimators are tractable and relevant for competition analysis. The two core applications, entry games with multiple equilibria and demand estimation without outside-good shares, address settings that competition economists encounter routinely. Both examples illustrate how weaker identifying assumptions can produce policy-relevant ranges rather than fragile point estimates.

The entry game application handles the multiplicity region in epsilon space, computes separate upper and lower probability bounds, and restricts attention to the competitive case $\boldsymbol{\delta_i < 0}$ for which the closed-form bounds are valid. The demand application treats the outside share as interval-identified, constructs unconditional moments from nonnegative basis functions of the cost shifter, and maps the resulting confidence set into merger counterfactuals through a logit-Bertrand simulation.

The main limitation is computational. For high-dimensional parameter spaces, grid search becomes infeasible, and practitioners must rely on projection methods, optimization-based approaches, adaptive grids, or parallel computation. A second limitation is interpretive: when bounds are wide, they may be uninformative for enforcement decisions. The practitioner must weigh the robustness advantage of weaker assumptions against the precision advantage of stronger ones.

The broader message is that partial identification is not a counsel of despair. Admitting that the data do not pin down a single answer is not a weakness. It is an honest accounting of what the evidence supports. When the bounds are tight enough to inform policy, and when they are robust to the assumptions that opposing experts would attack, moment inequality estimators provide a more credible foundation for competition analysis than point estimates built on fragile assumptions.

REFERENCES

Abito, J.M. and C. Chen (2023). A partial identification framework for dynamic games. *International Journal of Industrial Organization*, 87, Article 102915.

Andrews, D.W.K. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78 (1), 119–157.

Andrews, D.W.K. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica*, 81 (2), 609–666.

Beresteanu, A., I. Molchanov and F. Molinari (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79 (6), 1785–1821.

Berry, S. and E. Tamer (2006). Identification in models of oligopoly entry. In *Advances in Economics and Econometrics*, ed. R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press.

Bresnahan, T.F. (1982). The oligopoly solution concept is identified. *Economics Letters*, 10 (1-2), 87–92.

Bugni, F.A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78 (2), 735–753.

Bugni, F.A., I.A. Canay and X. Shi (2017). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8 (1), 1–38.

Bugni, F.A., J.L. Horowitz and L. Zhang (2026). Demand estimation without outside good shares. *Working Paper*, arXiv: 2602.19154.

Canay, I.A., G. Illanes and A. Velez (2024). A user's guide for inference in models defined by moment inequalities. *Journal of Econometrics*, 243 (1), Article 105558.

Chernozhukov, V., H. Hong and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75 (5), 1243–1284.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77 (6), 1791–1828.

Ciliberto, F., C. Murry and E. Tamer (2021). Market structure and competition in airline markets. *Journal of Political Economy*, 129 (11), 2995–3038.

Crawford, G. and A. Yurukoglu (2012). The welfare effects of bundling in multichannel television markets. *American Economic Review*, 102 (2), 643–685.

Dong, B., Y.-W. Hsieh and M. Shum (2021). Computing moment inequality models using constrained optimization. *Econometrics Journal*, 24 (3), 399–416.

Eizenberg, A. (2014). Upstream innovation and product variety in the U.S. home PC market. *Review of Economic Studies*, 81 (3), 1003–1045.

Ericson, R. and A. Pakes (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies*, 62 (1), 53–82.

Haile, P. and E. Tamer (2003). Inference with an incomplete model of English auctions. *Journal of Political Economy*, 111 (1), 1–51.

Ho, K. and A. Pakes (2014). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, 104 (12), 3841–3884.

Ho, K. and A. Rosen (2017). Partial identification in applied research: Benefits and challenges. In *Advances in Economics and Econometrics*, ed. B. Honoré, A. Pakes, M. Piazzesi and L. Samuelson. Cambridge: Cambridge University Press. NBER Working Paper No. 21641.

Holmes, T.J. (2011). The diffusion of Wal-Mart and economies of density. *Econometrica*, 79 (1), 253–302.

Kaido, H., F. Molinari and J. Stoye (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87 (4), 1397–1432.

Kline, B., A. Pakes and E. Tamer (2021). Moment inequalities and partial identification in industrial organization. In *Handbook of Industrial Organization*, Vol. 4, ed. K. Ho, A. Hortaçsu and A. Lizzeri. Amsterdam: Elsevier, 345–431.

Molinari, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics*, Vol. 7A, ed. S.N. Durlauf, L.P. Hansen, J.J. Heckman and R.L. Matzkin. Amsterdam: Elsevier, 355–486.

Pakes, A. (2010). Alternative models for moment inequalities. *Econometrica*, 78 (6), 1783–1822.

Pakes, A., J. Porter, K. Ho and J. Ishii (2015). Moment inequalities and their application. *Econometrica*, 83 (1), 315–334.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *Review of Economic Studies*, 70 (1), 147–165.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2, 167–195.

Wollmann, T.G. (2018). Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, 108 (6), 1364–1406.

Xiao, R. (2018). Identification and estimation of incomplete information games with multiple equilibria. *Journal of Econometrics*, 203 (2), 328–343.

Development of a Data-Driven Digital Twin for State of Health Estimation of Li-Ion Batteries

Hava Temel[1] and Furkan Soysal[2]

Abstract

The widespread adoption of electric vehicles and renewable energy systems has made accurate battery health monitoring a critical engineering challenge. This study proposes a data-driven Digital Twin framework for State of Health (SOH) estimation, End-of-Life (EOL) detection, and Remaining Useful Life (RUL) prediction of lithium-ion batteries using the NASA Prognostics Center of Excellence (PCoE) dataset. A five-criterion cleaning pipeline applied to 34 batteries across five operating groups yielded 18 quality-filtered batteries. Twenty-six cycle-level features covering voltage, current, temperature, impedance, and protocol information were extracted per discharge cycle. A global BiLSTM model with a 10-cycle sliding window trained on 14 batteries achieved MAPE = 2.85%, MAE = 0.0245, RMSE = 0.0331, and R^2 = 0.9323. A linearity test on SOH degradation curves showed linear regression was insufficient for 13 of 17 batteries (mean R^2 = 0.698), confirming non-linear degradation behavior and justifying the BiLSTM architecture. The Digital Twin layer translated SOH predictions into EOL detection and RUL estimation for all 18 batteries. In a generalization test on four completely unseen batteries from the most challenging operating group, the model successfully identified the critical EOL region. The proposed framework supports predictive maintenance, replacement planning, second-life assessment, and data-driven battery asset management.

Keywords: Battery Management System, Digital Twin, SOH Estimation, BiLSTM, Remaining Useful Life.

JEL Codes: N70, O13, Q40, C45, C63.

Lityum İyon Bataryaların Sağlık Durumu (SOH) Tahmini İçin Veri Güdümlü Bir Dijital İkiz Geliştirilmesi

Özet

Elektrikli araçların ve yenilenebilir enerji sistemlerinin yaygınlaşması, lityum-iyon bataryaların güvenli ve verimli kullanımını kritik bir mühendislik problemi haline getirmiştir. Bu sistemlerin performansı büyük ölçüde batarya sağlık durumuna (State of Health – SOH) bağlıdır. Ancak SOH doğrudan ölçülemediğinden dolaylı olarak tahmin edilmesi gerekmektedir. Geleneksel model tabanlı yaklaşımlar, batarya yaşlanmasının doğrusal olmayan ve zamana bağlı karmaşık doğasını yeterince temsil edememektedir. Bu çalışma, SOH kestirimi için veri odaklı bir Dijital İkiz yaklaşımı önermektedir. Önerilen çerçeve, fiziksel batarya sistemi ile sanal modeli arasında sürekli veri etkileşimi sağlayarak batarya davranışını dinamik olarak izlemeyi ve tahmin etmeyi amaçlamaktadır. Gerilim, akım, sıcaklık ve kapasite gibi zaman serisi verilerinin analizi, batarya yaşlanmasının doğrusal olmayan bir karaktere sahip olduğunu ve geçici kapasite toparlanmaları gibi karmaşık davranışlar içerdiğini göstermektedir. Geliştirilen Dijital İkiz mimarisi; veri toplama, ön işleme, modelleme ve senkronizasyon katmanlarından oluşmaktadır. Bu yapı sayesinde gerçek zamanlı izleme ve öngörücü analiz mümkün hale gelmekte, anormal durumların erken tespiti sağlanmaktadır. Böylece bakım maliyetlerinin azaltılması, sistem güvenliğinin artırılması ve batarya ömrünün daha etkin yönetilmesi hedeflenmektedir. Sonuç olarak, önerilen veri odaklı yaklaşım, batarya yönetim sistemleri için esnek ve ölçeklenebilir bir çözüm sunmakta olup enerji depolama, elektrikli ulaşım ve akıllı enerji sistemlerinde uygulanabilir önemli bir potansiyel taşımaktadır.

Anahtar Kelimeler: Batarya Yönetim Sistemi, Dijital İkiz, Sağlık Durumu (SOH) Tahmini, Yapay Zekâ, Elektrikli Araçlar.

[1] Hava Temel, Undergraduate Student, Department of Energy Systems Engineering, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
email: temelhava17@gmail.com, ORCID: 0009-0009-5073-7441.

[2] Furkan Soysal, Ph.D., Associate Professor, Department of Energy Systems Engineering, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
email: fsoysal@aybu.edu.tr, ORCID: 0000-0002-2558-2014.

JEL Kodları: N70, O13, Q40.

1. INTRODUCTION

The global transition toward sustainable energy systems has accelerated the electrification of transportation, with lithium-ion (Li-ion) batteries serving as the cornerstone technology in electric vehicles (EVs), grid-scale energy storage, and portable electronics. As battery energy density, safety, and longevity directly determine the performance and economic viability of these systems, accurate health monitoring has emerged as one of the most critical challenges in modern engineering (Barré et al., 2013; Xiong et al., 2018).

Battery degradation is driven by complex electrochemical mechanisms including capacity fade, internal resistance increase, lithium plating, and Solid Electrolyte Interphase (SEI) layer growth. These mechanisms are highly non-linear, stochastic, and sensitive to operating conditions such as temperature, charge/discharge rates, and depth of discharge (Berecibar et al., 2016; Feng et al., 2019). State of Health (SOH), defined as the ratio of current maximum available capacity to nominal capacity, is the primary metric for quantifying battery degradation:

$$SOH = \frac{Q_{current}}{Q_{nominal}} \times 100 \tag{1.1}$$

where $Q_{current}$ denotes the measured instantaneous discharge capacity and $Q_{nominal}$ the initial reference capacity. SOH cannot be measured directly — it must be estimated from observable signals including terminal voltage, current, temperature, and impedance (Plett, 2015; Hu et al., 2020).

Conventional Battery Management Systems (BMS) rely primarily on Equivalent Circuit Models (ECMs) for SOH estimation. Although computationally efficient, ECMs fail to capture the non-linear and time-varying aging dynamics of Li-ion batteries under heterogeneous real-world conditions (He et al., 2012). Data-driven approaches based on deep learning have emerged as powerful alternatives (Severson et al., 2019; Li et al., 2021). Among these, Bidirectional Long Short-Term Memory (BiLSTM) networks are particularly well-suited to SOH estimation, processing time-series in both temporal directions and achieving state-of-the-art accuracy (Schuster and Paliwal, 1997; Gao and Cheng, 2024).

Digital Twin (DT) technology further extends the data-driven paradigm by creating a dynamic virtual replica of the physical system synchronized with operational data, enabling EOL detection, RUL estimation, and maintenance decision support (Madni et al., 2019; Fuller et al., 2020; Qayyum et al., 2025). This study proposes an end-to-end Digital Twin framework with five principal contributions: (1) a five-criterion cycle-level data cleaning pipeline; (2) 26-feature extraction with statistical validation; (3) a global BiLSTM model trained across four operating groups; (4) a Digital Twin layer providing battery-specific EOL and RUL outputs; and (5) generalization testing on four completely unseen batteries.

2. THEORETICAL FRAMEWORK

Battery State of Health estimation sits at the intersection of electrochemical degradation theory and sequential machine learning. The theoretical basis of this study rests on two pillars: the physical mechanisms of battery aging and the computational principles of BiLSTM networks.

2.1. Electrochemical Degradation Mechanisms

Battery degradation is governed by well-established electrochemical phenomena. The SEI layer forms on the anode surface during initial cycling and continues to grow, consuming lithium ions and increasing ohmic resistance R_e. The evolution of R_e over time can be modeled as (Barré et al., 2013):

$$R_e(t) = R_{e,0} + \alpha \cdot t^{0.5} \tag{2.1}$$

where $R_{e,0}$ denotes the initial ohmic resistance, α the SEI growth coefficient, and t time. Charge transfer resistance R_{ct} also increases as the electrode-electrolyte interface degrades (Berecibar et al., 2016). These mechanisms collectively reduce the deliverable capacity, quantified as SOH. Critically, degradation is non-linear — progressing gradually in early cycles and accelerating near the End-of-Life threshold (Feng et al., 2019).

2.2. BiLSTM Network Architecture

BiLSTM networks maintain gated memory cells that enable long-range temporal dependency learning in degradation trajectories. The fundamental gate equations of an LSTM cell are (Hochreiter and Schmidhuber, 1997):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.3}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.4}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.5}$$

Where f_t denotes the forget gate, i_t the input gate, c_t the candidate cell state, and h_t the hidden state; σ denotes the sigmoid activation function and $\odot$ element-wise multiplication. The bidirectional extension processes each sequence in both forward and backward temporal directions (Schuster and Paliwal, 1997):

$$\overrightarrow{h_t} = \text{LSTM}(x_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = \text{LSTM}(x_t, \overleftarrow{h_{t+1}}) \tag{2.6}$$

$$y_t = [\overrightarrow{h_t}\,;\, \overleftarrow{h_t}] \tag{2.7}$$

2.3. Huber Loss Function

Huber Loss was selected for model training to limit the dominant influence of outlier discharge cycles on the gradient signal. Defined with threshold parameter δ, Huber Loss is expressed as:

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & | y - \hat{y} | \leq \delta \\ \delta \cdot | y - \hat{y} | - \frac{1}{2}\delta^2 & | y - \hat{y} | > \delta \end{cases} \tag{2.8}$$

where y denotes the true SOH value, ŷ the model prediction, and δ = 0.02 the threshold parameter. When the error remains below δ, Huber Loss applies a quadratic (MSE-like) penalty; for errors exceeding δ, it transitions to linear penalization, limiting the influence of outlier cycles on model training.

3. LITERATURE REVIEW

3.1. Model-Based SOH Estimation

Traditional BMS architectures rely on Equivalent Circuit Models (ECMs) and electrochemical models. ECMs offer computational efficiency but cannot adequately represent the highly non-linear aging dynamics of Li-ion batteries under varying conditions (He et al., 2012; Plett, 2015). Electrochemical models such as the Doyle-Fuller-Newman (DFN) model offer superior physical fidelity but impose

prohibitive computational costs and require precise internal parameters difficult to obtain non-destructively (Doyle et al., 1993). A common limitation of both approaches is parameter drift: as batteries age, model parameters shift away from calibrated values, requiring frequent recalibration that is impractical for real-time BMS applications (Berecibar et al., 2016).

3.2. Data-Driven and Deep Learning Approaches

LSTM-based models capture temporal dependencies in battery degradation sequences, achieving RMSE values below 3% on benchmark datasets (Li et al., 2021; Zhang et al., 2018). BiLSTM architectures consistently outperform unidirectional LSTM for SOH estimation tasks (Gao and Cheng, 2024; Yang et al., 2025), with attention-augmented variants achieving MAPE below 0.9% (Gao and Cheng, 2024). CNN-BiLSTM hybrid architectures further improve performance by combining spatial feature extraction with temporal modeling (Sun et al., 2022; Li et al., 2024). However, the majority of existing studies adopt per-battery model structures trained on a single battery or a single operating condition. Global model approaches covering diverse temperature and current protocols remain limited in the literature, and generalization testing on completely unseen operating groups is rarely applied.

3.3. Digital Twin Technology in Battery Management

Recent reviews confirm that LSTM and Transformer-based models are the most commonly deployed within Digital Twin architectures for battery health management (Qayyum et al., 2025; Zhao, 2024). Nair et al. (2024) demonstrated that DT-enabled predictions significantly outperform conventional approaches on the NASA dataset. A closer examination of existing BiLSTM-based studies reveals that while works such as Zhang et al. (2018) and Gao and Cheng (2024) achieve high prediction accuracy, they adopt single-group, per-battery model structures. This study aims to address this gap by proposing a comprehensive framework that integrates a global BiLSTM model trained across four operating groups, generalization testing on completely unseen batteries and Digital Twin EOL and RUL outputs.Furthermore, the applicability of Digital Twin architectures is rapidly expanding beyond terrestrial EVs into highly critical aerospace domains, as recently demonstrated in space battery management systems (Sbarra et al., 2025).

4. DATA AND METHODS

4.1. Dataset

The NASA Prognostics Center of Excellence (PCoE) battery aging dataset was used (Saha and Goebel, 2009; Chen, 2024). The dataset contains 18650 cylindrical Li-ion cells tested under controlled cycling conditions, with measurements of terminal voltage, current, temperature, discharged capacity, and EIS parameters including R_e (ohmic resistance) and R_{ct} (charge transfer resistance). Table 1 summarizes the five operating groups.

Table 1. NASA PCoE Battery Dataset Operating Groups

Group	Battery IDs	Temperature	Current Protocol	Role
reference	B0005, B0006, B0007, B0018	24°C	2A constant current	Training
mixed	B0033, B0034, B0036	24°C	Variable cut-off voltages	Training
low_1A	B0045, B0046, B0047, B0048	4°C	1A constant current	Training
low_2A	B0054, B0055, B0056	4°C	2A constant current	Training
low_multi	B0041, B0042, B0043, B0044	4°C	Mixed 1A and 4A	Unseen Test

4.2. Data Cleaning and Quality Filtering

A five-criterion cleaning pipeline was applied at discharge cycle level. First, SOH was constrained to the physically plausible range $0.30 \leq \text{SOH} \leq 1.05$. Second, capacity and discharge duration were required to be strictly positive. Third, cycles with capacity below 60% of each battery's own median capacity were flagged as low-capacity outliers — a battery-relative threshold superior to fixed absolute thresholds:

$$C_i < 0.60 \times C_{battery} \tag{4.1}$$

where C_i denotes the capacity of discharge cycle i and $\tilde{C}_{battery}$ the median capacity of the corresponding battery. Fourth, a minimum of 60 clean discharge cycles per battery was imposed to ensure sufficient temporal length for BiLSTM training. Fifth, batteries B0049–B0052 were excluded due to software crashes. Following these criteria, 18 of 34 batteries passed quality filtering. Figure 1 shows the capacity trajectories with detected outliers; Figure 2 presents the resulting battery selection.

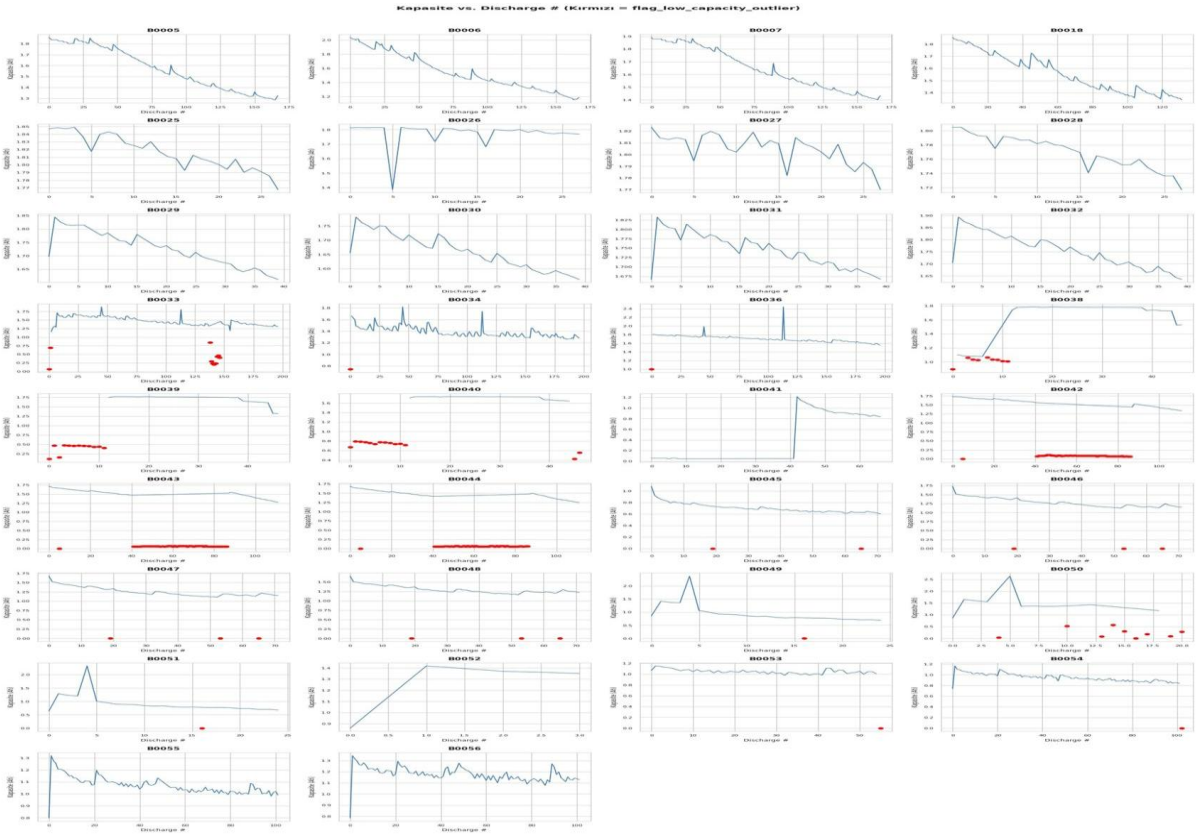


Figure 1. Capacity vs. Discharge Cycle Number for all 34 batteries.
Red markers indicate low-capacity outlier cycles flagged by the 60% battery-median criterion.
Source: Authors' calculations using NASA PCoE dataset.

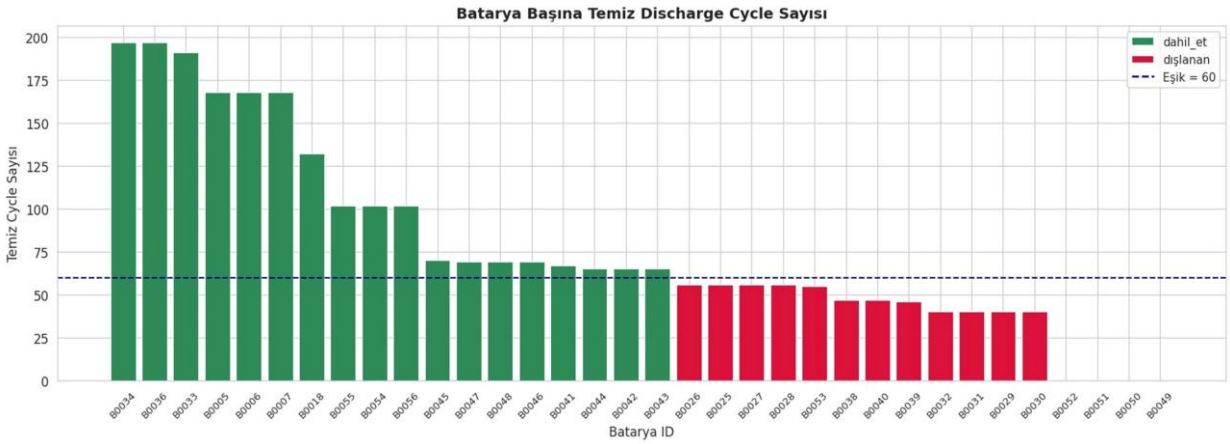


Figure 2. Clean discharge cycle count per battery after quality filtering.
Green: included (≥60 clean cycles). Red: excluded. Dashed line: 60-cycle threshold.
Source: Authors' calculations using NASA PCoE dataset.

4.3. SOH Calculation and Degradation Behavior

SOH was computed as the ratio of current discharge capacity to initial reference capacity. The 80% SOH threshold was adopted as the End-of-Life criterion (Barré et al., 2013):

$$SOH_{EOL} = 0.80 \tag{4.2}$$

$$RUL = N_{EOL} - N_{current} \tag{4.3}$$

where N_{EOL} denotes the predicted End-of-Life cycle and $N_{current}$ the current discharge cycle number. Figure 3 displays SOH trajectories for all 18 batteries, illustrating non-linear degradation, temporary recovery effects, and substantial variability across groups.

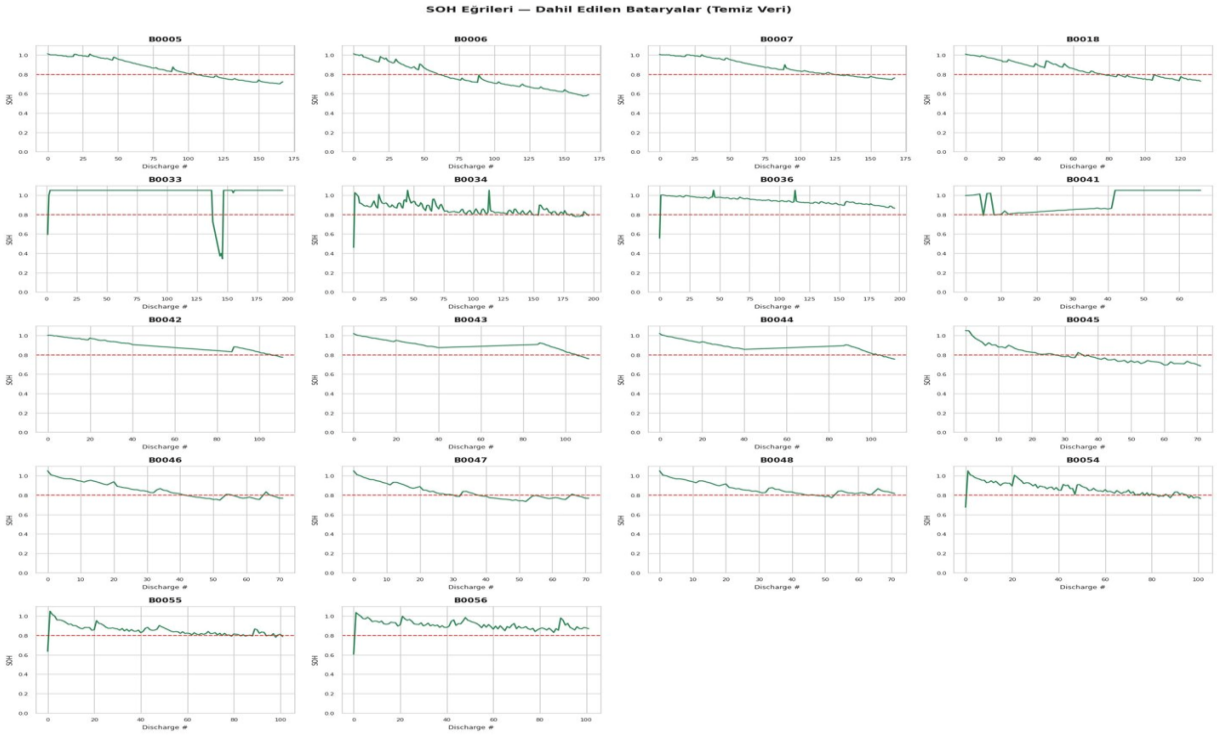


Figure 3. SOH trajectories of all 18 included batteries. Red dashed line: 80% EOL threshold.
Source: Authors' calculations using NASA PCoE dataset.

4.4. Feature Engineering and Statistical Validation

Twenty-six cycle-level features were computed from each discharge cycle (Table 2). To validate feature relevance, Pearson correlation coefficients between each feature and SOH were computed (Table 4.3). The discharge cycle number showed the strongest negative correlation with SOH ($r = -0.367$, $p < 0.001$), confirming that advancing cycles reduce battery health. Energy per cycle ($r = 0.336$), mean current ($r = 0.261$), and charge transfer resistance R_{ct} ($r = -0.153$) also showed statistically significant associations.

Importantly, ohmic resistance R_e yielded a near-zero and statistically insignificant correlation ($r = 0.001$, $p = 0.975$). This result reflects a known characteristic of the NASA PCoE dataset: EIS measurements are recorded at intervals rather than every discharge cycle, and missing values were forward-filled during preprocessing. Consequently, R_e values exhibit insufficient within-battery variance to produce a meaningful Pearson correlation. This constitutes a dataset-level limitation rather than a feature selection error — R_e remains physically motivated as a health indicator in electrochemical theory (Barré et al., 2013), and its inclusion preserves its potential utility in datasets with denser EIS recording. Global Min-Max normalization was applied jointly across training batteries; unseen batteries were transformed using the pre-fitted scaler without refitting, preventing data leakage.

Table 2. Feature Engineering: Categories, Feature Names, and Physical Relevance

Category	Feature Names	Count	Physical Relevance
Voltage	v_mean, v_min, v_max, v_std, v_drop_rate	5	Voltage drop rate reflects aging-induced impedance rise
Current	c_mean, c_std	2	Current variability affects degradation rate
Temperature	t_mean, t_max, ambient_temp_C	3	Higher temperature in aged batteries reflects increased internal resistance
Impedance	Re, Rct	2	Physical aging indicators (SEI growth, charge transfer degradation)
Time / Cycle	duration, energy_Wh, discharge_num, cutoff_V	4	Cycle energy and duration reflect capacity fade
Protocol (one-hot)	reference, mixed, low_1A, low_2A, low_multi	5	Encodes operating condition for cross-group generalization

Table 3. Pearson Correlation between Key Features and SOH (selected features)

Feature	Pearson r	p-value	Interpretation
discharge_num	−0.367	< 0.001	Strong: cycle count inversely tracks SOH
energy_Wh	0.336	< 0.001	Higher energy → healthier battery
c_mean	0.261	< 0.001	Moderate positive association
v_min	−0.205	< 0.001	Lower minimum voltage with aging
t_max	0.180	< 0.001	Temperature elevation with aging
Rct	−0.153	< 0.001	Charge transfer resistance grows with aging
Re	0.001	0.975	Not significant — EIS measurement sparsity*

* Re correlation is not statistically significant due to EIS measurement sparsity in the NASA PCoE dataset. Missing Re values were forward-filled, reducing within-battery variance. This is a dataset-level limitation, not a feature selection error.

4.5. Degradation Rate Analysis

To characterize aging behavior across operating groups, the per-cycle degradation rate was computed for each battery as the ratio of total SOH loss to total cycle count:

$$\lambda = \frac{SOH_{first} - SOH_{last}}{N_{cycle}} \tag{4.4}$$

where λ denotes the degradation rate, SOH_{first} and SOH_{last} the first and last observed SOH values, and

N_{cycle} the total number of clean discharge cycles. Batteries B0054, B0055, and B0056 (low_2A group) were excluded from this analysis as their initial SOH values fall below the 0.80 EOL threshold, indicating they entered the dataset in an already-degraded state. Battery B0041 was also excluded due to an anomalous increasing SOH profile. Table 4 presents group-level results.

The low_1A group exhibited the highest mean degradation rate (0.00410 SOH/cycle), substantially exceeding the reference group (0.00197 SOH/cycle). This finding is consistent with electrochemical theory: 4°C operating temperature reduces ion mobility and increases internal resistance, accelerating degradation relative to 24°C conditions (Berecibar et al., 2016). The mixed group showed the lowest degradation rate (0.00095 SOH/cycle), reflecting the lower depth-of-discharge induced by variable cut-off voltages.

Table 4. Group-Level Degradation Rate Analysis (mean SOH loss per discharge cycle)

Group	Batteries Included	Mean Degradation Rate (SOH/cycle)	Interpretation
reference	B0005, B0006, B0007, B0018	0.00197	Moderate, consistent aging at 24°C, 2A
mixed	B0034, B0036	0.00095	Slowest aging — variable cut-off reduces stress
low_1A	B0045, B0046, B0047, B0048	0.00410	Fastest aging — cold temperature accelerates degradation
low_multi	B0042, B0043, B0044	0.00226	High rate — cold temperature + variable current

Note: B0054, B0055, B0056 excluded (initial SOH below EOL threshold). B0041 excluded (anomalous increasing SOH profile). B0033 excluded (near-constant SOH throughout observation period).

4.6. SOH Linearity Test — Motivation for BiLSTM

To quantitatively justify the selection of a non-linear sequence model over simpler alternatives, a linear regression was fitted to the SOH trajectory of each battery, and the coefficient of determination R^2 was computed. Low R^2 indicates that linear models cannot adequately capture the degradation pattern, motivating the use of BiLSTM. Table 5 presents results. Thirteen of 17 batteries exhibit $R^2 < 0.90$, with a mean R^2 of 0.698, confirming predominantly non-linear degradation behavior. The reference group batteries (B0005, B0006, B0007, B0018) are notable exceptions with R^2 values of 0.94–0.98, reflecting their controlled 24°C, 2A conditions. The low_2A group shows particularly pronounced non-linearity (mean $R^2 = 0.45$), driven by irregular capacity fluctuations under cold temperature and higher current stress.

Table 5. SOH Linearity Test: R^2 of Linear Regression Fit per Battery

Battery	Group	R^2 (Linear Fit)	Non-Linear?
B0005	reference	0.976	No
B0006	reference	0.964	No
B0007	reference	0.976	No
B0018	reference	0.940	No
B0036	mixed	0.876	Yes
B0034	mixed	0.546	Yes
B0033	mixed	0.006	Yes
B0045	low_1A	0.857	Yes
B0046	low_1A	0.853	Yes
B0047	low_1A	0.755	Yes
B0048	low_1A	0.699	Yes
B0042	low_multi	0.953	No
B0043	low_multi	0.761	Yes
B0044	low_multi	0.769	Yes

B0054	low_2A	0.676	Yes
B0055	low_2A	0.471	Yes
B0056	low_2A	0.199	Yes
Mean / Total	—	0.698	13 / 17 non-linear

Threshold for linearity classification: $R^2 \geq 0.90$. B0041 excluded from analysis (anomalous SOH profile).

4.7. Global BiLSTM Model: Architecture and Training Configuration

The proposed framework consists of three interconnected pipeline stages (Figure 4): data preprocessing, global BiLSTM modeling, and Digital Twin output generation. A single global BiLSTM model was trained jointly on all 14 training batteries, pooling windows across heterogeneous operating conditions to learn shared aging dynamics. For each battery, a sliding window of 10 consecutive discharge cycles produced input tensors of shape (10, 26). The dataset was split 70%/15%/15% in cycle-wise temporal order.

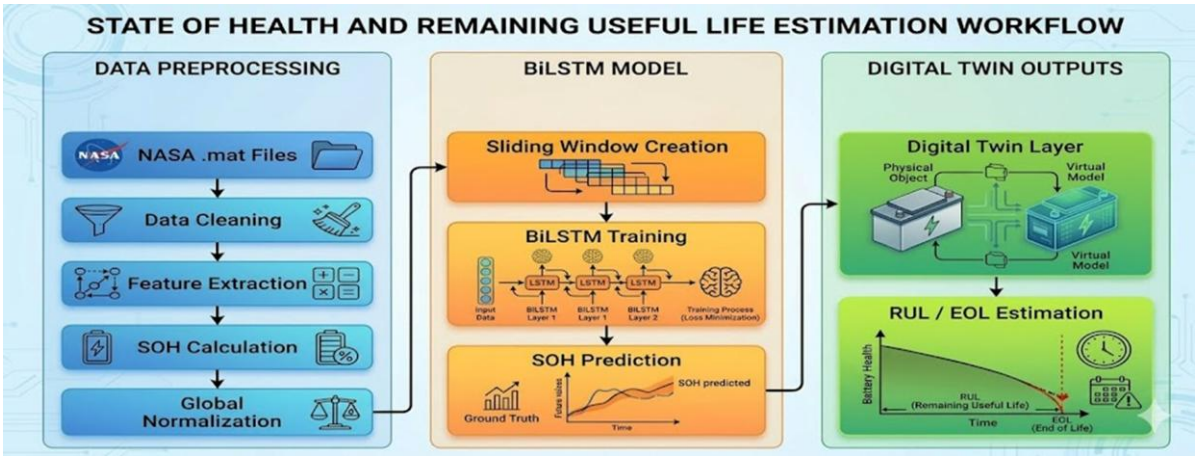


Figure 4. Proposed Data-Driven Digital Twin Framework.
Three stages: Data Preprocessing, Global BiLSTM Model, and Digital Twin Outputs.
Source: Figure generated using Google Gemini (AI image generation) based on the author's workflow design.
Note: This figure was generated with the assistance of Google Gemini (AI image generation tool) based on the workflow designed by the author, and is intended for illustrative purposes

Table 6. Global BiLSTM Model Architecture

Layer	Configuration	Purpose
Input	(None, 10, 26)	10-cycle sliding window; 26 features per cycle
BiLSTM — 1	64 units, return_sequences=True	Bidirectional temporal pattern extraction
Dropout — 1	Rate = 0.20	Regularization against overfitting
BiLSTM — 2	32 units	Sequential feature compression
Dropout — 2	Rate = 0.20	Second regularization layer
Dense	16 units, ReLU	Non-linear transformation toward SOH space
Output	1 unit, Linear	Continuous SOH regression output

Table 7. Training Configuration and Hyperparameters

Parameter	Value	Rationale
Window Size	10 cycles	Sufficient temporal context; avoids early-cycle data scarcity
Features	26 per cycle	Voltage, current, temperature, impedance, energy, protocol
Data Split	70% / 15% / 15%	Train / validation / test; cycle-wise temporal order preserved
Max Epochs	300	Upper bound; actual training halted at epoch 38
Batch Size	16	Balances memory efficiency and gradient stability
Optimizer	Adam (lr = 0.001)	Adaptive per-parameter learning rates
Loss Function	Huber Loss (δ = 0.02)	Robust to outlier discharge cycles
Early Stopping	Patience = 30 epochs	Halted at epoch 38; no overfitting observed
Training Batteries	14	Across 4 operating groups
Unseen Batteries	4 (B0041–B0044)	Withheld from all training and normalization fitting

Huber Loss (δ = 0.02) was selected over MSE to limit the gradient influence of outlier discharge cycles. Adam optimizer with learning rate 0.001 provides adaptive per-parameter updates suited to the multi-battery heterogeneous gradient landscape. Early stopping with patience=30 halted training at epoch 38, well before the 300-epoch maximum.

5. FINDINGS

5.1. Training Convergence

Figure 5 presents the training and validation Huber Loss curves across 38 epochs. Training loss decreased monotonically, confirming consistent learning. Validation loss exhibited moderate fluctuations attributable to the heterogeneous degradation behavior across operating groups — expected behavior for a global multi-battery model. No systematic divergence between training and validation loss was observed, confirming successful regularization. The final training Huber Loss converged to approximately 0.0003, while the validation loss stabilized around 0.0008, both well below the δ = 0.02 threshold.

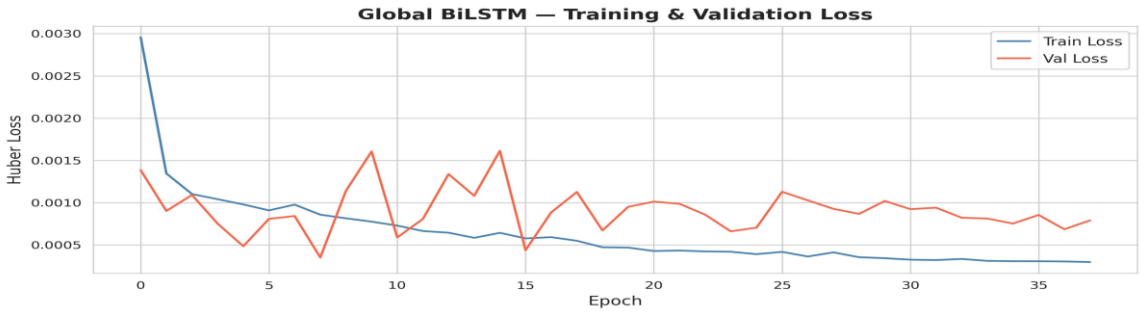


Figure 5. Global BiLSTM Training and Validation Loss (Huber Loss) across 38 epochs.*Early stopping halted training at epoch 38.*
Source: Authors' calculations using NASA PCoE dataset.

5.2. Model Performance

The global BiLSTM model achieved overall test performance of MAPE = 2.85%, MAE = 0.0245, RMSE = 0.0331, and R^2 = 0.9323 across 140 test sequences from 10 batteries. These results compare favorably with BiLSTM-based SOH estimators on the same NASA dataset (Zhang et al., 2018; Li et al., 2021). Zhang et al. (2018) reported RMSE values of 2–3% for LSTM models on the NASA dataset; the present global BiLSTM achieves comparable accuracy while training across four heterogeneous operating groups simultaneously. Figure 6. presents per-battery MAPE values; Table 8 provides full

performance metrics.

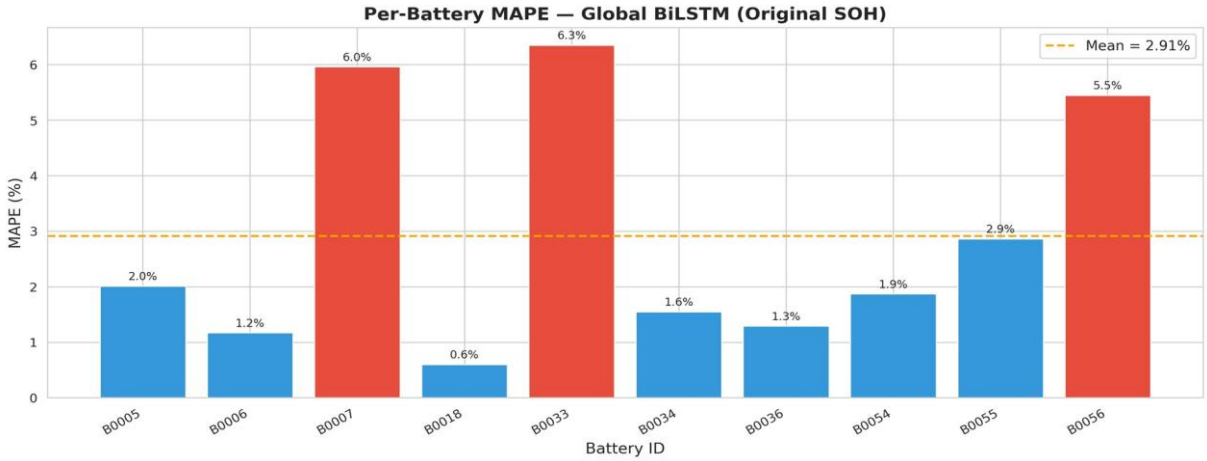


Figure 6. Per-Battery MAPE (%) for the Global BiLSTM Model.
Blue: below mean (2.91%). Red: above mean. Dashed line: mean MAPE = 2.91%.
Source: Authors' calculations using NASA PCoE dataset.

Table 8. Per-Battery Model Performance on Held-Out Test Set

Battery	Group	MAE	RMSE	MAPE (%)	R^2	Notes
B0005	reference	0.0143	0.0172	2.01	−5.56*	—
B0006	reference	0.0069	0.0078	1.17	0.658	EOL reached
B0007	reference	0.0451	0.0459	5.96	—†	—
B0018	reference	0.0045	0.0064	0.60	0.600	—
B0033	mixed	0.0666	0.0668	6.35	—†	Near-constant SOH
B0034	mixed	0.0124	0.0154	1.55	−0.077	—
B0036	mixed	0.0114	0.0126	1.29	−0.574	RUL active
B0054	low_2A	0.0144	0.0160	1.87	−1.831	—
B0055	low_2A	0.0227	0.0260	2.86	−3.047	—
B0056	low_2A	0.0478	0.0483	5.45	—†	RUL = 32 cycles
Overall / Mean	—	0.0245	0.0331	2.85	0.9323	n = 140 sequences

* Negative R^2 values arise when prediction error exceeds variance of the target variable within the test window — a known artifact when the test SOH range is very narrow. † R^2 flagged as unreliable due to near-zero test variance. In all such cases, MAE and MAPE confirm low absolute prediction errors. Overall R^2 = 0.9323 computed across all 140 test sequences.

Reference group batteries (B0018: MAPE = 0.60%, B0006: 1.17%) achieved the lowest errors, reflecting consistent degradation under standard conditions. Batteries with irregular SOH behavior — B0033 (6.35%) and B0007 (5.96%) — showed higher errors, consistent with the modeling difficulty of non-monotonic or cold-temperature degradation trajectories.

5.3. Digital Twin: EOL Detection and RUL Estimation

The Digital Twin layer processes the model's SOH trajectory to detect EOL and estimate RUL for each battery. EOL is defined as the first cycle at which predicted SOH falls below 0.80. Two scenarios are handled: EOL_predicted (battery has passed the threshold) and EOL_extrapolated (SOH above 0.80, EOL projected by linear trend extrapolation). Table 9 presents complete Digital Twin results for all 18 batteries.

Table 9. Digital Twin EOL Detection and RUL Estimation — All 18 Batteries

Battery	Group	Current SOH	Total Cycles	Pred. EOL Cycle	Est. RUL	Status
B0005	reference	0.722	167	111	0	EOL reached
B0006	reference	0.589	167	80	0	EOL reached
B0007	reference	0.762	167	130	0	EOL reached
B0018	reference	0.729	131	94	0	EOL reached
B0033	mixed	1.050	196	437	241	EOL extrapolated
B0034	mixed	0.790	196	184	0	EOL reached
B0036	mixed	0.866	196	291	95	EOL extrapolated
B0045	low_1A	0.686	71	25	0	EOL reached
B0046	low_1A	0.768	71	46	0	EOL reached
B0047	low_1A	0.767	71	49	0	EOL reached
B0048	low_1A	0.816	71	50	0	EOL reached
B0054	low_2A	0.764	101	88	0	EOL reached
B0055	low_2A	0.788	101	123	22	EOL extrapolated
B0056	low_2A	0.870	101	133	32	EOL extrapolated
B0041	low_multi (unseen)	1.050	66	10	0	Anomalous start SOH
B0042	low_multi (unseen)	0.774	111	94	0	EOL reached
B0043	low_multi (unseen)	0.759	111	95	0	EOL reached
B0044	low_multi (unseen)	0.755	111	90	0	EOL reached

Figures 7–9 illustrate Digital Twin outputs for three representative training batteries spanning different lifecycle stages. B0006 (Figure 7) has clearly passed the EOL threshold (current SOH = 0.589, predicted EOL at cycle 80, RUL = 0). B0036 (Figure 8) remains healthy at SOH = 0.866 with RUL = 95 cycles. B0056 (Figure 9) represents a near-EOL case under cold-temperature conditions, with RUL = 32 cycles — providing actionable near-term maintenance planning information.

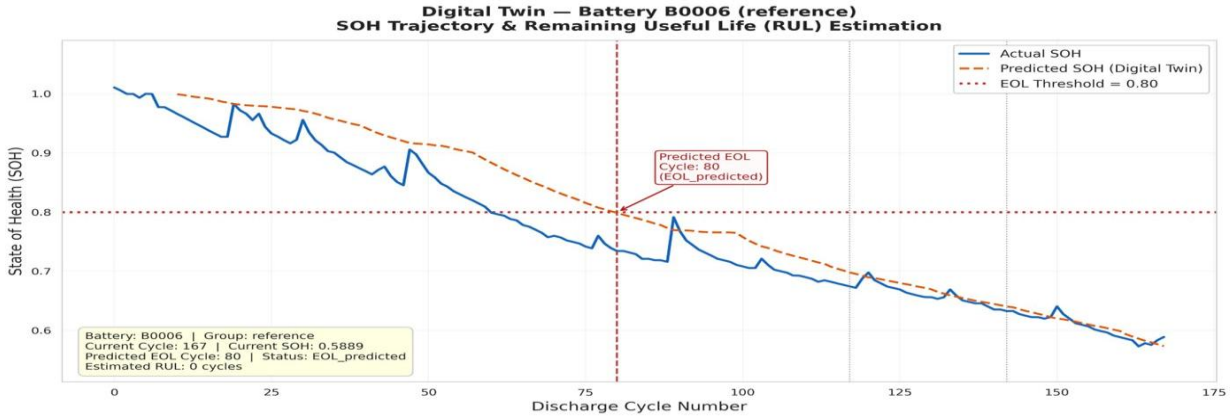


Figure 7. Digital Twin output for B0006 (reference group).
Current SOH = 0.589. Predicted EOL: cycle 80. Estimated RUL: 0 cycles (EOL_predicted).
Source: Authors' calculations using NASA PCoE dataset.

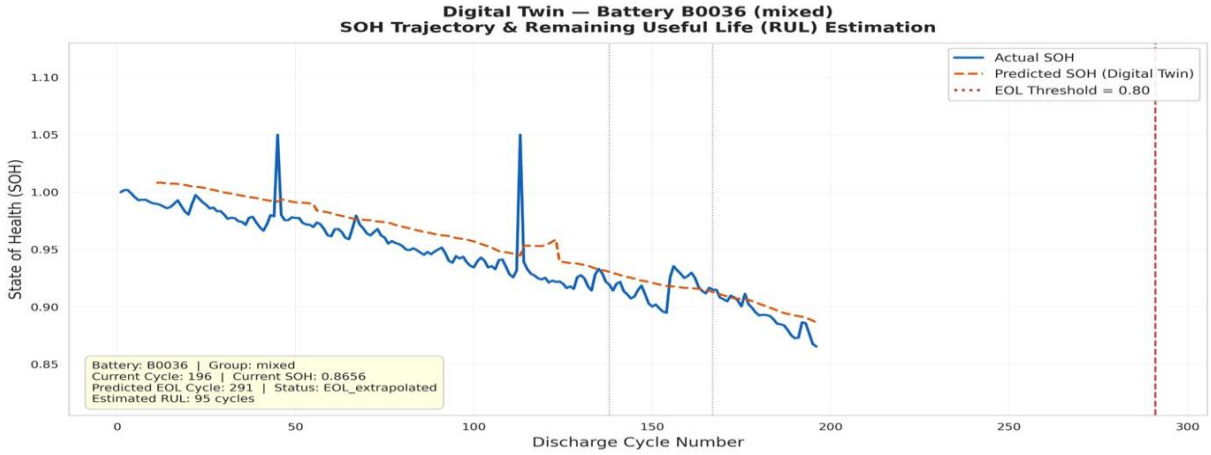


Figure 8. Digital Twin output for B0036 (mixed group).
Current SOH = 0.866. Predicted EOL: cycle 291. Estimated RUL: 95 cycles (EOL_extrapolated).
Source: Authors' calculations using NASA PCoE dataset.

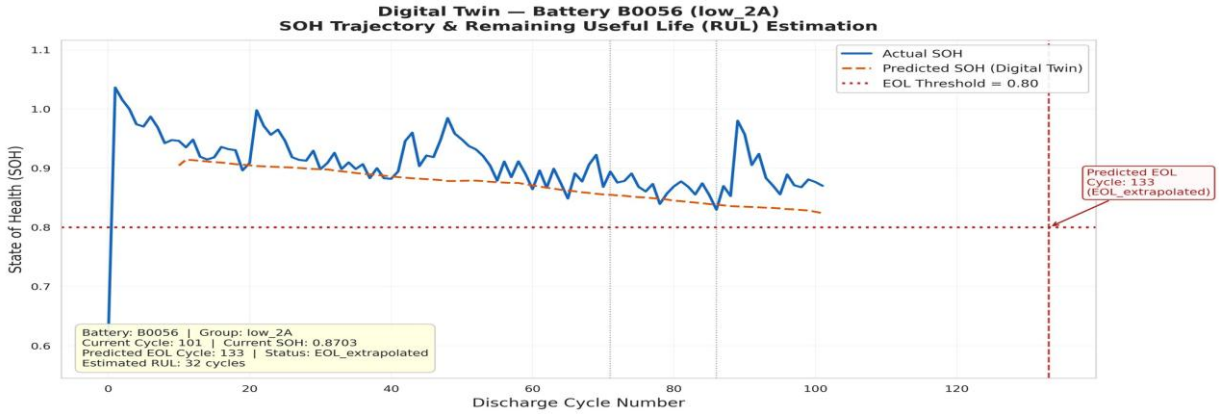


Figure 9. Digital Twin output for B0056 (low_2A group).
Current SOH = 0.870. Predicted EOL: cycle 133. Estimated RUL: 32 cycles (EOL_extrapolated).
Source: Authors' calculations using NASA PCoE dataset.

5.4. Unseen Battery Generalization Test

Figure 10 presents the Digital Twin output for B0042 (low_multi group, 4°C, mixed 1A/4A), the primary generalization test. This battery was excluded from all training, validation, and normalization fitting. The model predicted EOL at cycle 94; the actual SOH reached the 0.80 threshold at approximately the same point — confirming successful EOL region identification despite a MAPE of 6.45%. The mean MAPE across all four unseen batteries (B0041–B0044) was 7.12%, reflecting the domain shift from the training distribution. Battery B0041 showed an anomalous increasing SOH profile with initial SOH above 1.0, which is attributed to measurement artifacts in the dataset rather than model failure.

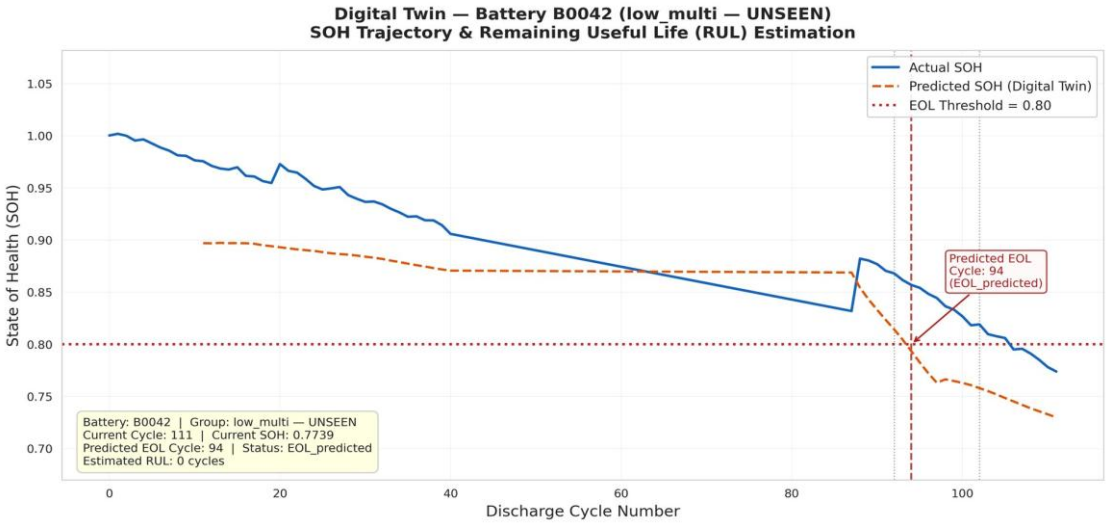


Figure 10. Digital Twin output for B0042 (low_multi — UNSEEN).
Current SOH = 0.774. Predicted EOL: cycle 94. Estimated RUL: 0 cycles. EOL region correctly identified despite domain shift.
Source: Authors' calculations using NASA PCoE dataset.

This result demonstrates initial-level EOL region detection capability for unseen batteries under challenging domain shift. Binary EOL status determination — the most operationally critical output — is achieved correctly for B0042, B0043, and B0044. These findings suggest that the global BiLSTM framework holds promise for cross-condition deployment, while highlighting the need for more diverse training data to improve trajectory-level accuracy under domain shift.

5.5. Study Limitations

This study has several limitations that should be considered when interpreting the results. First, the proposed framework was developed and evaluated exclusively using the NASA Prognostics Center of Excellence (PCoE) battery aging dataset. While this dataset is widely recognized and frequently used in battery health research, it may not fully capture the diversity of operating conditions encountered in real-world applications. Second, the analysis was performed using laboratory-generated data rather than operational Battery Management System (BMS) data collected from electric vehicles or stationary energy storage systems. Consequently, the model's real-world performance remains to be validated. Third, the study focused solely on a BiLSTM-based architecture and did not include a direct comparison with alternative deep learning approaches such as Transformer-based, attention-enhanced, or hybrid neural network models. Finally, although the proposed framework demonstrated promising generalization capability, prediction accuracy decreased for completely unseen operating conditions, indicating that broader training datasets covering a wider range of temperatures, current profiles, and degradation behaviors are necessary for robust deployment in practical applications. Despite these limitations, the results provide a strong foundation for the development of data-driven Digital Twin frameworks for battery health monitoring and predictive maintenance.These limitations also provide opportunities for future research and further validation of Digital Twin frameworks under real-world operating conditions.

6. CONCLUSION AND POLICY IMPLICATIONS

This study developed and validated a complete end-to-end data-driven Digital Twin framework for lithium-ion battery State of Health (SOH) estimation, End-of-Life (EOL) detection, and Remaining Useful Life (RUL) estimation across five heterogeneous operating groups using the NASA PCoE dataset. A systematic data-cleaning pipeline reduced 34 batteries to 18 quality-filtered batteries suitable for analysis. Statistical validation identified discharge cycle number ($r = -0.367$), energy per cycle ($r =$

0.336), and charge transfer resistance R_{ct} (r = −0.153) as the most informative predictors of battery health, while the insignificant correlation of Re was attributed to EIS measurement sparsity within the dataset. Degradation-rate analysis revealed substantial differences among operating conditions, with the low_1A group exhibiting the fastest degradation (0.00410 SOH/cycle) and the mixed group the slowest (0.00095 SOH/cycle). Furthermore, the linearity analysis demonstrated that 13 of 17 batteries followed predominantly non-linear degradation trajectories (mean $R^2 = 0.698$), providing quantitative justification for the selection of a BiLSTM-based architecture.

To the best of the authors' knowledge, this study is among the first to integrate a global BiLSTM model trained across multiple operating conditions with explicit generalization testing on completely unseen batteries and a functional Digital Twin layer capable of generating battery-specific EOL and RUL outputs. The proposed model achieved strong predictive performance, with MAPE = 2.85%, MAE = 0.0245, RMSE = 0.0331, and $R^2 = 0.9323$. In addition, the Digital Twin framework successfully generated EOL and RUL estimates for all 18 batteries and correctly identified the binary EOL status for three of four completely unseen batteries, demonstrating promising cross-condition deployment capability.

Beyond its technical contributions, the proposed framework carries important implications for battery asset management, energy policy, and the future deployment of electric vehicles and energy storage systems. The results support the integration of predictive battery health monitoring into next-generation Battery Management Systems (BMS), enabling proactive maintenance, optimized replacement planning, and more reliable second-life battery assessment. The study also highlights the importance of comprehensive battery-aging datasets, as the performance reduction observed in unseen operating conditions underscores the need for broader and more diverse real-world battery data. From a policy perspective, encouraging standardized battery data collection and supporting open-access degradation datasets could accelerate the development of robust and universally deployable battery health management technologies. Ultimately, the transition from reactive maintenance strategies toward data-driven predictive asset management has the potential to improve battery safety, reduce lifecycle costs, and enhance the economic viability of electric mobility and grid-scale energy storage systems.

Future research should focus on uncertainty quantification through Bayesian LSTM or Monte Carlo Dropout approaches, comparisons with Transformer-based and attention-enhanced architectures, validation using real-world BMS datasets, and direct RUL regression modeling as an alternative to SOH trajectory extrapolation.

Table 10. Digital Twin Outputs and Supported Decision Processes

Application	Digital Twin Output	Decision Enabled
Predictive Maintenance	RUL estimation	Schedule replacement before failure; reduce emergency costs
Replacement Planning	EOL detection + RUL	Individual-level battery swap scheduling
Second-Life Assessment	SOH trajectory	Identify batteries suitable for stationary storage re-use
Fleet Asset Management	SOH monitoring	Continuous health tracking across battery fleets
Safety Management	Early EOL detection	Proactive intervention before safety-critical failure modes
Data-Driven Operations	Full pipeline output	Replace rule-based BMS with model-based management

REFERENCES

Barré, A., B. Deguilhem, S. Grolleau, M. Gérard, F. Suard, and D. Riu (2013). A review on lithium-ion battery ageing mechanisms and estimations for automotive applications. *Journal of Power Sources*, 241, 680–689.

Berecibar, M., I. Gandiaga, I. Villarreal, N. Omar, J. Van Mierlo, and P. Van den Bossche (2016). Critical review of state of health estimation methods of Li-ion batteries for real applications. *Renewable and Sustainable Energy Reviews*, 56, 572–587.

Chen, Y. (2024). NASA lithium-ion battery dataset. *IEEE DataPort*. https://ieee-dataport.org/documents/nasa-lithium-ion-battery-dataset

Doyle, M., T. F. Fuller, and J. Newman (1993). Modeling of galvanostatic charge and discharge of the lithium/polymer/insertion cell. *Journal of the Electrochemical Society*, 140(6), 1526–1533.

Feng, X., X. He, M. Ouyang, L. Lu, D. Ren, and S. Santhanagopalan (2019). Thermal runaway mechanism of lithium-ion battery for electric vehicles: A review. *Energy Storage Materials*, 10, 246–267.

Fuller, A., Z. Fan, C. Day, and C. Barlow (2020). Digital twin: Enabling technologies, challenges and open research. *IEEE Access*, 8, 108952–108971.

Gao, D. and Y. Cheng (2024). Application of state of health estimation and remaining useful life prediction for lithium-ion batteries based on AT-CNN-BiLSTM. *Scientific Reports*, 14, 29026.

He, H., R. Xiong, and J. Fan (2012). Evaluation of lithium-ion battery equivalent circuit models for state of charge estimation by an experimental approach. *Energies*, 5(12), 4762–4781.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hu, X., L. Xu, X. Lin, and M. Pecht (2020). Battery lifetime prognostics. *Joule*, 4(2), 310–346.

Li, Y., K. Liu, A. M. Foley, A. Zülke, M. Berecibar, E. Nanini-Maury, J. Van Mierlo, and H. E. Hoster (2021). Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. *Renewable and Sustainable Energy Reviews*, 113, 109254.

Li, Z., X. Zhang, and W. Gao (2024). State of health estimation of lithium-ion battery during fast charging process based on BiLSTM-Transformer. *Energy*, 311, 133418.

Madni, A. M., C. C. Madni, and S. D. Lucero (2019). Leveraging digital twin technology in model-based systems engineering. *Systems*, 7(1), 7.

Nair, A. K. et al. (2024). AI-driven digital twin model for reliable lithium-ion battery discharge capacity predictions. *International Journal of Intelligent Systems*, 2024, 8185044.

Plett, G. L. (2015). Battery Management Systems, Volume I: Battery Modeling. Norwood, MA: *Artech House*.

Qayyum, A. et al. (2025). Digital twins for battery health prognosis: A comprehensive review of recent advances and challenges. *Journal of Power Sources*.

Saha, B. and K. Goebel (2009). Battery Data Set. *NASA Ames Prognostics Data Repository*, NASA Ames Research Center, Moffett Field, CA.

Sbarra, R. G., M. Pasquali, G. Coppotelli, P. Gaudenzi, and D. di Ienno et al. (2025). Digital twins for space battery management systems. *Energies*, 18(21), 5858.

Schuster, M. and K. K. Paliwal (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.

Severson, K. A., P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, and R. D. Braatz (2019). Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5), 383–391.

Sun, S., J. Sun, Z. Wang et al. (2022). Prediction of battery SOH by CNN-BiLSTM network fused with attention mechanism. *Energies*, 15(12), 4428.

Xiong, R., J. Cao, Q. Yu, H. He, and F. Sun (2018). Critical review on the battery state of charge estimation methods for electric vehicles. *IEEE Access*, 6, 1832–1843.

Yang, M., Y. Liu, B. Li, and C. Yang (2025). State of health estimation for lithium-ion batteries with an attention-integrated BiLSTM-MLP hybrid model. *Communications in Computer and Information Science*, 2420. Springer.

Zhang, Y., R. Xiong, H. He, and M. G. Pecht (2018). Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. *IEEE Transactions on Vehicular Technology*, 67(7), 5695–5705.

Zhao, K. (2024). Digital twin battery storage in energy storage system. PhD Thesis, *Cardiff University*.

Artificial Intelligence in Finance and FinTech: Applications, Risks, and Regulatory Frontiers

Ibrahim Musa Ünal[1], **Fatma Nur Aysan**[2] **and Ahmet Faruk Aysan**[3]

Abstract

This chapter examines the growing role of artificial intelligence (AI) in finance and the FinTech ecosystem, with particular attention to its applications in risk management, financial forecasting, and credit scoring. It explores how machine learning techniques and data-driven models are transforming the way financial institutions evaluate risk, predict market dynamics, and assess borrower creditworthiness. The chapter then analyzes the rise of AI-enabled algorithmic trading and its implications for market efficiency, liquidity, and price discovery. At the same time, it discusses the regulatory challenges and systemic risk considerations that emerge from the widespread adoption of AI in financial markets, including issues related to model transparency, data bias, operational resilience, and market stability. By evaluating both the opportunities and the potential risks associated with AI-driven financial innovation, the chapter provides a balanced perspective on the evolving intersection between artificial intelligence, financial markets, and regulatory governance. The analysis draws on academic literature, regulatory reports, and recent industry developments to argue that the productive integration of AI in finance depends not only on algorithmic sophistication but also on the strength of governance frameworks, explainability standards, and supervisory capacity.

Keywords: Artificial Intelligence, FinTech, Machine Learning, Credit Scoring, Algorithmic Trading, Financial Regulation.

JEL Codes: G17, G21, G28, O33, C45.

Finans ve FinTech'te Yapay Zekâ: Uygulamalar, Riskler ve Düzenleme Alanındaki Gelişmeler

Özet

Bu bölüm, yapay zekânın (AI) finans ve FinTek ekosistemindeki giderek artan rolünü incelemeyi amaçlamakta; özellikle risk yönetimi, finansal tahminleme ve kredi skorlama alanlarındaki uygulamalarına odaklanmaktadır. Makine öğrenmesi teknikleri ve veri odaklı modellerin, finansal kurumların riski değerlendirme, piyasa dinamiklerini öngörme ve borçluların kredi değerliliğini analiz etme biçimlerini nasıl dönüştürdüğünü ele almaktadır. Bölüm ayrıca, yapay zekâ destekli algoritmik işlemlerin yükselişini ve bunun piyasa etkinliği, likidite ve fiyat keşfi üzerindeki etkilerini analiz etmektedir. Bununla birlikte, finansal piyasalarda yapay zekânın yaygın kullanımından doğan düzenleyici zorluklar ve sistemik risk unsurları da tartışılmaktadır. Bu kapsamda, model şeffaflığı, veri yanlılığı, operasyonel dayanıklılık ve piyasa istikrarı gibi konulara dikkat çekilmektedir. Yapay zekâ temelli finansal inovasyonun sunduğu fırsatlar ile potansiyel riskleri birlikte değerlendiren bu bölüm, yapay zekâ, finansal piyasalar ve düzenleyici yönetişim arasındaki gelişen etkileşime dengeli bir bakış açısı sunmayı hedeflemektedir.

Anahtar Kelimeler: Yapay Zekâ, Fintek, Risk Yönetimi.

JEL Kodları: C45, G21, C32.

[1] Ibrahim Musa Ünal, Qatar Financial Markets Authority – Department of Risk, Doha, Qatar
email: i.unal@qfma.org.qa, ORCID: 0000-0001-6802-8446.
[2] Fatma Nur Aysan, Istanbul University, İstanbul, Türkiye
email: fatimanur@gmail.com, ORCID: 0009-0008-6267-0495.
[3] Ahmet Faruk Aysan, Hamad Bin Khalifa University – Department of Islamic Finance and Economy, Doha, Qatar
email: aaysan@hbku.edu.qa, ORCID: 0000-0001-7363-0116.

1. INTRODUCTION

Artificial intelligence (AI) has moved from the periphery of financial innovation to the center of how modern financial services are designed, delivered, and supervised. Over the last decade, advances in machine learning, the availability of granular and high-frequency data, and substantial increases in computing capacity have combined to make AI a general-purpose technology for finance (Aldasoro et al., 2024; Financial Stability Board [FSB], 2017). Banks, asset managers, insurers, market infrastructures, and FinTech firms now rely on AI techniques for tasks ranging from anti-money-laundering surveillance and credit underwriting to portfolio construction, execution, and customer interaction.

The economic logic for this shift is well established. AI systems excel at pattern recognition in large, heterogeneous datasets, at modeling non-linear relationships, and at adapting to changing data distributions through retraining (Goodfellow et al., 2016; Heaton et al., 2017). These capabilities map naturally onto core financial problems, which typically involve forecasting uncertain outcomes, screening signals from noisy data, and managing high-dimensional risks. At the same time, the rise of FinTech has produced a class of entrants whose business models are essentially AI-native, in which decisions about pricing, eligibility, and risk are made primarily by algorithms (Philippon, 2016; Frost et al., 2019).

Yet this expansion of AI in finance is not without tension. Supervisors and academic researchers have raised concerns about model opacity, data bias, procyclicality, and the potential for new forms of systemic risk arising from common models, shared datasets, and concentration in cloud and AI service providers (Danielsson et al., 2022; Bank for International Settlements [BIS], 2024; FSB, 2024). Recent episodes of market stress and the global debate around generative AI have intensified scrutiny on how AI systems behave under tail conditions and how their use should be governed (Acemoglu, 2024). The objective of this chapter is to map this evolving landscape, focusing on three application areas — risk management, financial forecasting, and credit scoring — and on two cross-cutting issues — AI-driven trading and the regulatory frontier.

The remainder of the chapter is structured as follows. Section 2 introduces the foundations of AI in finance and FinTech and clarifies key concepts. Section 3 examines AI applications in risk management. Section 4 turns to financial forecasting. Section 5 analyzes AI in credit scoring. Section 6 considers algorithmic and AI-enabled trading. Section 7 discusses regulatory challenges and systemic risk considerations. Section 8 concludes.

2. FOUNDATIONS OF AI IN FINANCE AND FINTECH

2.1. From Statistical Models to Machine Learning

Quantitative finance has long relied on statistical and econometric models. What distinguishes contemporary AI applications is the systematic use of flexible, high-capacity learners — including regularized regressions, tree-based ensembles, gradient-boosted models, support vector machines, and deep neural networks — that can exploit large feature spaces without strong distributional assumptions (Hastie et al., 2009; Goodfellow et al., 2016). Gu et al. (2020) provide a benchmark study showing that machine learning methods substantially improve out-of-sample prediction of asset returns relative to classical linear models, primarily by capturing non-linear interactions among predictors.

FinTech, in turn, refers to the broader set of technology-enabled financial innovations in payments, lending, asset management, insurance, and market infrastructures (Philippon, 2016; Thakor, 2020). AI is not synonymous with FinTech, but it is increasingly the analytical engine inside FinTech products. As Frost et al. (2019) observe, the competitive edge of large platforms in financial services derives less from their balance sheets than from their data and algorithmic capacity.

2.2. Data, Models, and Infrastructure

Three complementary inputs shape the deployment of AI in finance. The first is data: alongside traditional financial records, institutions now ingest transaction streams, alternative data such as mobile telephony and e-commerce footprints, text from disclosures and news, and increasingly multimodal sources (Berg et al., 2020; Bartlett et al., 2022). The second is models: from logistic regression at one end to large language models at the other. The third is infrastructure, including cloud platforms, specialized hardware, and feature stores, which determines how quickly models can be trained, monitored, and retired (BIS, 2024).

Aldasoro et al. (2024) emphasize that the AI stack in finance is becoming increasingly modular, with foundation models supplied by a small number of external providers and adapted by financial firms for downstream tasks. This structure offers efficiency gains but also concentrates dependencies, an issue revisited in Section 7.

3. AI IN FINANCIAL RISK MANAGEMENT

Risk management is one of the most mature application areas for AI in finance. Banks, insurers, and securities firms use machine learning across credit, market, operational, liquidity, and conduct risks. The common motivation is the same: replace or augment models that summarize complex realities with linear approximations by models that can absorb more information and adapt more quickly to changing conditions (Bank of England, 2022).

3.1. Market and Credit Risk Modeling

In market risk, machine learning techniques have been used to improve value-at-risk and expected shortfall estimation, to forecast volatility, and to identify regime changes. Bucci (2020) shows that neural networks can outperform classical GARCH-family models in volatility forecasting for several major indices, particularly at higher frequencies. In credit risk, ensemble methods such as random forests and gradient boosting consistently improve discrimination over logistic regression when borrower-level features are rich and non-linear effects are present (Khandani et al., 2010; Bracke et al., 2019).

These gains, however, depend critically on data quality and on the stability of the underlying environment. Bracke et al. (2019) caution that more accurate models can also be more opaque and more sensitive to distributional shifts, which raises questions about validation, stress testing, and the assignment of responsibility when models fail.

3.2. Operational Risk, Fraud, and Anti-Money Laundering

AI is also used extensively in operational risk, fraud detection, and anti-money-laundering (AML) compliance. Unsupervised and semi-supervised methods help identify anomalous transactions, while graph-based approaches uncover hidden relationships in payment networks (Weber et al., 2019). These tools have proven valuable in environments where labeled examples are scarce and adversaries adapt their behavior. The Financial Action Task Force (FATF, 2021) notes that AI can substantially improve the precision of transaction-monitoring systems but also warns that poor model governance can entrench false positives and disproportionately affect certain customer segments.

3.3. Model Risk and Governance

As AI models proliferate, model risk itself becomes a first-order concern. Supervisory expectations have long required firms to manage the life cycle of models, including documentation, validation, ongoing monitoring, and challenge functions (Board of Governors of the Federal Reserve System [Fed], 2011). The European Banking Authority (EBA, 2021) and the Bank of England (2022) have explicitly extended these expectations to machine learning models, calling for proportionality, interpretability where

decisions affect customers, and clear human accountability.

4. AI AND FINANCIAL FORECASTING

4.1. Return, Volatility, and Macro-Financial Forecasting

Forecasting future prices, returns, and volatility is among the oldest problems in finance. The empirical literature has historically been skeptical of out-of-sample predictability, but recent contributions using machine learning suggest that modest but economically meaningful improvements are possible, particularly when many weak signals are combined (Gu et al., 2020; Chen et al., 2024). Deep learning architectures such as long short-term memory networks have been applied to sequence prediction in equity and foreign-exchange markets with some success, although results are sensitive to the choice of evaluation period and to transaction costs (Sezer et al., 2020).

At the macro-financial level, machine learning models are increasingly used to nowcast economic activity from heterogeneous indicators, including search trends, payments data, and satellite imagery (Bok et al., 2018). Central banks have begun to integrate these tools into their analytical toolkit while maintaining traditional structural models as anchors for policy interpretation (BIS, 2024).

4.2. Text, Sentiment, and Large Language Models

A second strand of forecasting research applies natural language processing to corporate disclosures, regulatory filings, news, and social media. Loughran and McDonald (2011) provide a widely used dictionary for finance-specific sentiment analysis, and subsequent work has shown that text-based features carry information about future returns, volatility, and credit events beyond standard accounting variables. Large language models extend this frontier by allowing more flexible extraction of meaning, including the identification of forward-looking statements and risk factors (Lopez-Lira and Tang, 2024). Their use in production, however, raises concerns about hallucination, prompt sensitivity, and reproducibility, which remain active areas of research.

5. AI IN CREDIT SCORING AND LENDING

5.1. From Traditional Scorecards to Machine Learning Underwriting

Credit scoring has been one of the earliest and most consequential applications of AI in retail finance. Traditional scorecards built on logistic regression remain widely used because of their simplicity and interpretability, but a growing body of evidence shows that machine learning models, particularly gradient-boosted trees and neural networks, deliver higher discriminatory power across a range of consumer credit settings (Khandani et al., 2010; Albanesi and Vamossy, 2019).

In emerging markets and among thin-file borrowers, the inclusion of alternative data has been transformative. Berg et al. (2020) demonstrate that digital footprint variables — such as device type, time of application, and email provider — have predictive power for default that is comparable to traditional credit bureau scores. Frost et al. (2019) and Gambacorta et al. (2023) document that platform-based lenders combining AI models with alternative data can reach borrowers who are underserved by incumbent banks, contributing to financial inclusion in several jurisdictions.

5.2. Fairness, Bias, and Explainability

These efficiency and inclusion gains are accompanied by significant fairness and accountability questions. Bartlett et al. (2022) provide evidence that algorithmic lenders in the United States discriminated less than face-to-face lenders on average but still produced disparate pricing outcomes for minority borrowers. The mechanism is not necessarily intent: models trained on historical data can reproduce or amplify pre-existing inequalities through correlated proxies (Barocas and Selbst, 2016;

Fuster et al., 2022). Fuster et al. (2022) show that more complex models can shift the distribution of credit access across demographic groups in ways that are not visible from aggregate accuracy metrics alone.

Regulators have responded by emphasizing explainability and the ability of consumers to contest adverse decisions (European Commission, 2024; Consumer Financial Protection Bureau [CFPB], 2023). Post-hoc explanation tools such as SHAP and LIME have become standard components of model risk management, though their limitations as faithful representations of model behavior are increasingly recognized in the academic literature (Rudin, 2019).

6. AI-ENABLED ALGORITHMIC TRADING AND MARKET MICROSTRUCTURE

6.1. From Execution Algorithms to AI Traders

Algorithmic trading is now responsible for a large share of activity in equities, futures, and foreign-exchange markets (Securities and Exchange Commission [SEC], 2020). Execution algorithms designed to minimize implementation shortfall, smart order routers, and market-making algorithms have used statistical and rule-based methods for many years. The recent frontier is the explicit use of reinforcement learning and deep learning to design adaptive trading strategies (Buehler et al., 2019). Buehler et al. (2019) introduce the notion of deep hedging, in which neural networks learn hedging policies under transaction costs and market frictions directly from simulated or historical data.

6.2. Implications for Liquidity, Efficiency, and Stability

The literature on algorithmic trading is broadly consistent in finding that, under normal conditions, it improves liquidity and price efficiency, especially for liquid instruments (Hendershott et al., 2011). Under stressed conditions, however, the picture is more nuanced. Kirilenko et al. (2017) show that high-frequency traders did not cause the 2010 Flash Crash but contributed to its dynamics through inventory management. More recent analyses raise the possibility that AI traders pursuing similar reward functions and trained on similar data could converge on highly correlated behavior, increasing the likelihood of liquidity dry-ups and short-lived crashes (Danielsson et al., 2022).

The question of whether AI-driven traders can learn collusive or destabilizing strategies in continuous markets is the subject of growing research interest. Calvano et al. (2020) show, in a different context, that simple reinforcement learning agents can learn tacit collusion in repeated pricing games, a finding whose extension to financial markets remains an open empirical question.

7. REGULATORY CHALLENGES AND SYSTEMIC RISK CONSIDERATIONS

7.1. Model Transparency and Explainability

The first cluster of regulatory concerns relates to the opacity of advanced AI models. Supervisors require that decisions affecting consumers — particularly in credit, insurance pricing, and certain investment recommendations — be explainable and contestable (EBA, 2021; CFPB, 2023). The European Union's AI Act (European Commission, 2024) classifies a number of financial use cases, including credit scoring, as high-risk and imposes requirements on documentation, data governance, human oversight, and transparency. Rudin (2019) argues forcefully that for high-stakes decisions, regulators and firms should prefer models that are interpretable by design rather than relying solely on post-hoc explanations of black-box models.

7.2. Data Bias and Consumer Protection

A second cluster concerns data bias and its consequences for consumer protection. AI systems trained on historical decisions can encode patterns of discrimination even when sensitive attributes are excluded

from the inputs (Barocas and Selbst, 2016). The CFPB (2023) has clarified that adverse-action notice requirements under the Equal Credit Opportunity Act apply to AI-based credit decisions, and similar expectations are emerging in other jurisdictions. These rules effectively bind the design space of acceptable models: explainability and the ability to articulate the principal reasons for a decision are not optional features but legal obligations.

7.3. Operational Resilience and Third-Party Dependencies

A third cluster of concerns is operational resilience. As AI components are increasingly delivered through external providers — including cloud platforms and foundation-model vendors — financial institutions face new forms of third-party risk concentration (BIS, 2024; FSB, 2024). The European Union's Digital Operational Resilience Act addresses some of these risks for critical information and communication technology providers in finance, including testing, incident reporting, and oversight regimes. Aldasoro et al. (2024) caution that supervisory frameworks designed for institution-level risk may be poorly suited to systems where a small number of upstream providers shape outcomes across the financial system.

7.4. Systemic Risk and Market Stability

Finally, the systemic implications of AI in finance are receiving increasing attention. Danielsson et al. (2022) argue that widespread adoption of similar AI models can introduce new sources of procyclicality and herding, particularly when models are calibrated on overlapping data and optimize similar objectives. The FSB (2024) highlights cross-border coordination, data governance, and supervisory capacity as priority issues, while the BIS (2024) emphasizes the need for central banks to invest in their own AI capabilities to monitor and respond to AI-related developments in real time. These themes are likely to define the regulatory agenda for the remainder of the decade.

8. CONCLUSION

This chapter has surveyed the role of AI in finance and the FinTech ecosystem, with attention to risk management, financial forecasting, credit scoring, algorithmic trading, and the regulatory frontier. The evidence supports a measured optimism. AI techniques deliver measurable gains in predictive accuracy across a wide range of financial tasks; they can extend access to credit and services to underserved populations; and they offer supervisors and central banks new tools for monitoring complex systems. At the same time, the same techniques can entrench bias, opacity, and concentration, and they can interact with market microstructure in ways that are not yet fully understood.

Three implications follow. First, the productive integration of AI in finance is as much a governance problem as a technical problem, and progress in one without the other is unlikely to be durable. Second, the explainability and accountability of AI systems should be treated as engineering requirements from the outset rather than as compliance overlays added at the end. Third, the regulatory community will need to deepen its own technical capacity and to coordinate across jurisdictions to address the cross-border and cross-sectoral nature of AI risks. The trajectory of AI in finance over the next decade will depend less on the raw capability of models than on the institutional choices made now about data, transparency, and accountability.

REFERENCES

Acemoglu, D. (2024). The simple macroeconomics of AI. *Economics Policy*, 40(121), 1–45.

Albanesi, S., and Vamossy, D. F. (2019). Predicting consumer default: A deep learning approach. *NBER Working Paper* No. 26165. National Bureau of Economic Research, Cambridge, MA.

Aldasoro, I., Doerr, S., Gambacorta, L., Notra, S., Oliviero, T., and Whyte, D. (2024). Generative

artificial intelligence and cyber security in central banking. *BIS Papers* No. 145. Bank for International Settlements, Basel.

Bank for International Settlements. (2024). *Artificial intelligence and the economy: Implications for central banks. BIS Annual Economic Report 2024*. Basel: BIS.

Bank of England. (2022). *Machine learning in UK financial services*. London: Bank of England and Financial Conduct Authority.

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732.

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 143(1), 30–56.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *Review of Financial Studies*, 33(7), 2845–2897.

Board of Governors of the Federal Reserve System. (2011). *Supervisory guidance on model risk management (SR 11-7)*. Washington, DC: Federal Reserve.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10, 615–643.

Bracke, P., Datta, A., Jung, C., and Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. *Bank of England Staff Working Paper* No. 816. London: Bank of England.

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3), 502–531.

Buehler, H., Gonon, L., Teichmann, J., and Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8), 1271–1291.

Calvano, E., Calzolari, G., Denicolò, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2), 714–750.

Consumer Financial Protection Bureau. (2023). *Consumer financial protection circular 2023-03: Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B*. Washington, DC: CFPB.

Danielsson, J., Macrae, R., and Uthemann, A. (2022). Artificial intelligence and systemic risk. *Journal of Banking and Finance*, 140, 106290.

European Banking Authority. (2021). *EBA discussion paper on machine learning for IRB models*. Paris: EBA.

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Brussels: European Commission.

Financial Action Task Force. (2021). *Opportunities and challenges of new technologies for AML/CFT*. Paris: FATF.

Financial Stability Board. (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. Basel: FSB.

Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. Basel: FSB.

Frost, J., Gambacorta, L., Huang, Y., Shin, H. S., and Zbinden, P. (2019). BigTech and the changing structure of financial intermediation. *Economic Policy*, 34(100), 761–799.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.

Gambacorta, L., Huang, Y., Qiu, H., and Wang, J. (2023). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm. *Journal of Financial Intermediation*, 56, 101055.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1), 3–12.

Hendershott, T., Jones, C. M., and Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *Journal of Finance*, 66(1), 1–33.

Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11), 2767–2787.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The Flash Crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3), 967–998.

Lopez-Lira, A., and Tang, Y. (2024). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working paper*. University of Florida.

Loughran, T., and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Philippon, T. (2016). The FinTech opportunity. *NBER Working Paper* No. 22476. National Bureau of Economic Research, Cambridge, MA.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Securities and Exchange Commission. (2020). *Staff report on algorithmic trading in U.S. capital markets*. Washington, DC: SEC.

Sezer, O. B., Gudelek, M. U., and Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.

Thakor, A. V. (2020). Fintech and banking: What do we know? *Journal of Financial Intermediation*, 41, 100833.

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., and Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *KDD Workshop on Anomaly Detection in Finance*. Anchorage, AK.

DATAMACLEA'26

SPONSORLAR | SPONSORS		
INFOYATIRIM		
TCMB	TURKISH AIRLINES	ODTÜ - MEBİVA 1995
TÜBİTAK ULAKBİM	TÜBİTAK BİLGEM	
IEREK GmbH	YZTD 2020 YAPAY ZEKA VE TEKNOLOJİ DERNEĞİ	ADAM Ankara Düşünce ve Araştırma Merkezi
DESTEKLEYEN ÜNİVERSİTELER | PARTNER UNIVERSITIES		
Ankara Yıldırım Beyazıt Üniversitesi	Sivas Cumhuriyet Üniversitesi 1974	Necmettin Erbakan Üniversitesi
T.C. İstanbul Kültür Üniversitesi	Fatih Sultan Mehmet Vakıf Üniversitesi 2010	Ankara Yıldırım Beyazıt Üniversitesi
		İşletme ve Ekonomi Kulübü