Advanced AI Models for Future Forecasting of Budget Expenditures via Machine Learning and Deep Learning

Yunus Emre Gür
Department of Management Information Systems, Firat University, 23119 Merkez, Elazig, Türkiye E-mail: yegur@firat.edu.tr
ORCID: 0000-0001-6530-0598

Abdunnur Yıldız
Department of Finance, Firat University, 23119 Merkez, Elazig, Türkiye
E-mail: abdunnur@firat.edu.tr
ORCID: 0000-0002-6068-3363

Emre Ünal (Corresponding author)
Department of Economics, Firat University, 23119 Merkez, Elazig, Türkiye
E-mail: eunal@firat.edu.tr
ORCID: 0000-0001-9572-8923

Received: 29 September 2024, Accepted: 8 July 2025

Summary: This study builds on Türkiye's long-standing challenges in managing public spending amid political and economic uncertainty. Budget planning plays a vital role in ensuring fiscal sustainability and economic resilience. Therefore, this study proposes a new forecasting framework that utilizes the latest artificial intelligence models. This paper aims to provide data-driven decision support to policymakers by improving the accuracy and robustness of expenditure forecasts under complex temporal dynamics. Accordingly, the use of machine learning and deep learning methods to forecast budget expenditures in Türkiye was proposed and analyzed. Comprehensive datasets extending from January 2008 to May 2024 was considered. Datasets also include various extraordinary periods such as the global financial crisis, the European debt crisis, various major political events in Türkiye, and the COVID-19 pandemic. Model performance was evaluated using the Time Fusion Transform, which has received praised for its superior performance even in complex and volatile time series. The model results show that the MAPE is 0.0658%, MAE is 0.0050%, RMSE is 0.0111%, and R^2 is 0.993%. The Random Search learning algorithm was implemented to determine the optimal hyperparameters that enable the model to work effectively on the data. According to the findings of the study, the model can perform well despite economic and political changes. Biodiversity in the use of all model types shows that machine learning and deep learning models also offer valuable insights into the budget forecasting process.

Keywords: Budget Forecasting, Deep Learning, Machine Learning, Public Finance, Temporal Fusion Transformer.

JEL Classifications: C45, C53, C61, H68.

The persistent volatility in Türkiye's public finance system caused by political uncertainties, macroeconomic shocks and external crises has increased the need for more adaptive and robust forecasting tools. Traditional methods often fail to capture the non-linear and dynamic behavior of budget trends. Recent advances in artificial intelligence (AI), especially deep learning-based time series models such as Temporal Fusion Transformer (TFT), offer a new opportunity to overcome these challenges. Against this backdrop, the objective of this study is to investigate the effectiveness of machine learning (ML) and deep learning (DL) approaches in forecasting Türkiye's budget expenditures and assesses their potential to support more accurate and resilient fiscal planning. This study evaluates the differences between various models in terms of their ability to reliably forecast budget expenditures using historical data and aims to provide insights into how these advanced methodologies can improve fiscal forecasts and inform government planning processes. This study contributes to bridging this gap by not only comparing individual AI models, but also by building new ensemble and hybrid architectures to assess their performance in a volatile fiscal environment. Based on this framework, this paper addresses the following open research questions: How effective are advanced ML and DL models in forecasting Türkiye's public budget expenditures? Which of these models exhibits superior forecasting performance in this context? What practical policy implications can be derived from comparative forecasting results? These questions form the basis for the analyses conducted and guide the evaluation of AI-based forecasting strategies in the public finance context. In addition to its practical relevance for national fiscal planning, this research contributes to the broader literature by showing how AI-based models, especially those that can handle non-linear, high-dimensional time series data, can outperform traditional econometric techniques in complex public finance settings. The findings provide empirical evidence supporting a methodological shift towards data-driven forecasting in the public sector and provide a transferable framework for other emerging economies facing similar fiscal volatility.

Advances in data analytics and AI have transformed economic and financial forecasting processes; ML and DL models have improved forecasting accuracy by extracting meaningful patterns from complex data structures and surpassing traditional methods (Aliu, 2019; Lin and Huang, 2020; Maeda et al., 2021; Hossain et al., 2022). These models provide powerful tools for forecasting financial trends and decision support processes (Zhou et al., 2024). Furthermore, the tax system, which is a fundamental elements of public finance, supports economic growth and welfare by ensuring a balance between income and expenditure. In this context, timely and accurate forecasting of budget expenditures is vital for fiscal planning (Obadić et al., 2014; Kairu et al., 2021). While reliable forecasts facilitate efficient public resource allocation, poor forecasts can jeopardize fiscal sustainability (Robinson, 1998).

The rest of the paper is organized as follows: Section 1 reviews related literature and defines the research gap. Section 2 describes the datasets, preprocessing, and models used. Section 3 presents empirical results and model evaluations. Section 4 concludes with key findings and policy implications.

1. Previous research and the current work

Accurate forecasting of public budget expenditures plays a central role in ensuring fiscal discipline, resource efficiency, and long-term macroeconomic stability. Well-founded expenditure predictions are essential for aligning fiscal planning with national development strategies, thereby enabling governments to allocate limited resources more effectively (Devarajan et al., 1996; Zatonatska et al., 2022). This becomes particularly critical during crises or post-disaster recovery when targeted fiscal responses are required to minimize economic disruption (Sun, 2015). Reliable forecasting systems help to preemptively address macroeconomic imbalances such as current account deficits (Šuliková et al., 2014; Dzigbede et al., 2022; Özcan and Günal, 2024), and underpin multi-annual budgeting frameworks while reinforcing prudent fiscal management (Shkolnyk et al., 2021). Additionally, optimizing state budget structures can enhance both national growth outcomes and the state's ability to meet social needs when paired with predictive analytics (Kuzheliev et al., 2019).

Although the global literature has increasingly embraced AI-based techniques to improve forecasting accuracy, Türkiye's budget system has long struggled with structural and political complexities that have undermined fiscal predictability. The main aim of this study is to investigate whether new-generation AI Technologies, particularly ML and DL models, can enhance the accuracy and robustness of Türkiye's budget expenditure forecasts. Beyond addressing a national need, this research also aims to identify the forecasting models best suited to the Turkish context. Türkiye was selected as the case study because of data availability and its structurally volatile fiscal environment. Over the past few decades, the country has faced repeated macroeconomic shocks, high inflation cycles, frequent fiscal revisions, and inconsistent public spending patterns, all of which create a uniquely challenging context for budgetary forecasting. These conditions increase the non-linearity and uncertainty of expenditure dynamics, making Türkiye a compelling and high-risk testing ground for evaluating the robustness of advanced forecasting models. Furthermore, the persistent forecast deviations documented in prior empirical studies (e.g., Özcan, 2017; Kara, 2024a) underscore the need to explore more adaptive methodologies in this setting. Beyond its fiscal dynamics, Türkiye continues to struggle with a chronic current account deficit that has persisted for decades despite the implementation of various growth-oriented policy measures. This structural imbalance not only exacerbates the country's vulnerability to external shocks and political uncertainties but also fuels long-standing macroeconomic instabilities, such as persistent exchange rate depreciation, subdued economic growth, recurrent trade deficits, and elevated default risk (Köse and Ünal, 2024). These interconnected risks further complicate fiscal management and economic forecasting in the context of Türkiye. Thus, the country provides a compelling case for examining AI-based fiscal forecasting in emerging economies, particularly considering the persistent challenges in Türkiye's budget management.

The findings of this study are expected to have meaningful policy implications for politicians, economists, investors, and practitioners involved in fiscal planning. Moreover, this study seeks to offer insights that can be adapted for use in other developing nations facing similar fiscal constraints. In this context, this study evaluates an end-to-end forecasting framework, including data collection, preprocessing techniques, and comparative model performance. Prior studies on Türkiye's public finance landscape have explored a range of issues, including healthcare funding, corruption, financial crises, political instability, government overspending, and climate-related expenditures (Yardım et al., 2013; Eryılmaz and Murat, 2016; Çınaroğlu and Başer, 2019; Özekicioğlu and Tülümce, 2020; Akkaya, 2022; Naumoski et al., 2022; Önal, 2024). However, none of these works have employed ML or DL models for expenditure forecasting. This study addresses this research gap by not only applying AI models such as eXtreme Gradient Boosting (XGBoost), Multi Layer Perceptron (MLP), Long Short-Term Memory (LSTM) and Random Forest, but also introducing the TFT, a cutting-edge, interpretable DL model capable of capturing complex temporal dependencies. The ability of TFT to handle both static and dynamic covariates over multi-horizon forecasts offers substantial advantages for modeling fiscal dynamics. Furthermore, this study extends beyond individual model applications by incorporating two ensemble approaches: one based on a Voting Regressor combining ML models (random forest, XGBoost, and MLP), and another based on manually integrating TFT and LSTM predictions. Two hybrid models are constructed by extracting high-level features from TFT and LSTM networks and using them as inputs for an ML-based ensemble model. This layered architecture leverages the strengths of both deep and traditional learners to enable a more comprehensive evaluation of forecasting performance. A comparative analysis of these models reveals the transformative potential of AI applications in public finance and presents a novel contribution to the growing body of interdisciplinary research in this field.

1.1 Previous research

The shift from traditional statistical models to advanced AI-based approaches has reshaped the field of financial forecasting. From a theoretical perspective, traditional econometric forecasting methods, such as linear regression or ARIMA, assume certain functional forms, linearity, and stationary data that are often invalid in real-world financial environments. Furthermore, these assumptions are often invalid in the complex and dynamic environment of real-world finance, making traditional methods less effective in capturing the nuanced behavior of financial markets (Zakaria et al., 2023; Olubusola et al., 2024; Ajiga et al., 2024). The primary limitation of traditional models is their inability to capture non-linear relationships or sudden structural breaks in financial data, which can severely skew forecasts and lead to misguided decisions (Olubusola et al., 2024). In contrast, ML and AI methodologies offer a more flexible, data-driven approach to modeling. These advanced techniques can adapt to the inherent complexities and high-dimensional structure of financial data, making them particularly useful for uncovering hidden patterns that traditional models may miss. By leveraging comprehensive datasets, AI models can process and analyze changes in financial indices, budget expenditures, and market trends more efficiently than traditional statistical methods (Ajiga et al., 2024; Kanupriya, 2024). The data-driven nature of these AI models allows for the inclusion of numerous variables and interactions in the data, unlike classical models, which are limited to linear representations. Moreover, the advent of ML has ushered in an era of improved forecasting accuracy and precision. Recent research has highlighted the significant strides that AI-driven financial forecasting has made in improving forecasting capabilities, especially in market trend analysis and asset price prediction, using methods such as DL and reinforcement learning (Olubusola et al., 2024; Ajiga et al., 2024). These methodologies can identify complex interdependencies and regime changes inherent in financial data that traditional econometric models struggle to capture. Therefore, integrating AI methods can revolutionize sectors such as finance and budget management by providing precise forecasts that account for non-linearities and dynamic changes in market behavior. The flexibility of AI techniques is particularly advantageous in environments characterized by volatility and sudden changes, which are common in financial markets. For instance, traditional models may struggle to provide timely forecasts during market disruptions caused by economic changes or geopolitical events, whereas AI algorithms can pivot based on new data insights and effectively recalibrate forecasts as financial conditions evolve (Ajiga et al., 2024). This adaptability is crucial for financial institutions and agencies tasked with budget management because it allows for agile responses to emerging economic scenarios and further validates AI's superiority over traditional forecasting models for financial applications (Zakaria et al., 2023; Olubusola et al., 2024; Ajiga et al., 2024). The vital role of data preprocessing in determining model performance is among the most emphasized themes in the literatüre. Empirical evidence suggests that appropriate transformations, such as logarithmic or inverse hyperbolic sine (IHS), can significantly enhance the forecasting accuracy, sometimes even more so than the model selection itself. Comparison between classical models (e.g., ARIMA and exponential smoothing) and ML models, such as K-nearest neighbors (KNN) and neural networks, in the sales tax forecasting context have demonstrated such findings (Larson and Overton, 2024).

Another recurring theme in the literature is the increasing use of interpretable DL architectures for time series forecasting. For example, TFT has emerged as a powerful model for handling long-range dependencies and multi-horizon forecasting. It allows for static and dynamic variables to be integrated while maintaining interpretability, which is essential for decision-makers. TFT's capabilities have been validated in various domains, including GDP forecasting (Laborda et al., 2023), tourism demand during the COVID-19 pandemic (Wu et al., 2022), and complex economic systems that require long-term projections (Lim et al., 2019; Yun et al., 2023). Moreover, feature reduction and hybrid model combinations have gained traction as strategies for improving forecasting robustness. PCA-based dimensionality reduction, followed by supervised learning models such as support vector regression (SVR), has shown promise in public finance contexts. For instance, fiscal revenues in Henan Province were effectively forecasted using a PCA-SVR model, which outperformed classical time-series models in handling seasonal and structural shocks (Yu, 2024). Similarly, gated recurrent unit (GRU) networks have demonstrated strong performance in predicting government expenditures using historical macroeconomic data, outperforming a range of models including ARIMA, LSTM, SVR, and XGBoost (Yang et al., 2023).

The application of ML models to simulate and optimize public budgets also represents an important direction. Valle-Cruz et al. (2022) explored how Random Forest and synthetic data can be used for budget modeling in Mexico, concluding that such approaches provide better support for fiscal decision-making than deterministic rules. Similarly, Noor et al. (2022) addressed the forecasting of Malaysia's federal revenues and found that traditional linear regression models yielded high residual errors, whereas feedforward neural networks (FFNN) and Random Forest models produced far more accurate results. Finally, ensemble learning techniques have been shown to be beneficial for forecasting at the municipal level. Studies applying neural networks and support vector machines in ensemble configurations using methods such as bagging, boosting, and rotation forests have demonstrated improved forecasting performance over linear models, especially when relevant socioeconomic indicators, such as population, enterprise count, and tax base are included (Hájek and Olej, 2010).

Collectively, this body of research indicates that ML and DL approaches, particularly when combined with sound feature engineering and data preprocessing, have substantial potential for enhancing the accuracy, adaptability, and interpretability of budget forecasts across various governmental contexts. However, despite global progress, a clear gap remains in the application of these techniques to Türkiye's public finance system.

1.2 The current work and the research gap

Although a considerable body of literature has examined budget forecasting in Türkiye, the dominant methodological approach remains grounded in traditional econometric models. These studies have largely focused on identifying patterns of deviation in forecast accuracy over time, without incorporating more modern, data-driven predictive frameworks. Forecast underestimations have been consistently reported in various contexts, including education expenditures and tax revenue estimations. These deviations have also been observed across multiple timeframes and are typically attributed to institutional or structural limitations rather than inflationary effects or macroeconomic variables (Şenesen, 2000; Bağdigen, 2002; Yılmaz, 2003).

Temporal analyses covering longer historical periods have consistently revealed systematic weaknesses in Türkiye's budget forecasting system. In particular, year-end appropriation estimates tend to outperform initial forecasts, indicating inefficiencies in budget planning. This pattern has been observed across both general budgetary frameworks and sector-specific contexts, including education financing (Özcan and Tosun, 2014; Özcan, 2017). Although institutional reforms were introduced after 2006 with the aim of enhancing forecasting stability the macroeconomic repercussions of poor forecasting, such as inflationary pressure and interest rate volatility, remain prevalent concerns (Parlak, 2005; Yaşa et al., 2020). Recent findings have reinforced this thematic consensus by demonstrating that forecast deviations persist even in the post-reform era. An analysis of Türkiye's medium-term budget performance between 2009 and 2023 revealed significant inaccuracies in both revenue and expenditure forecasts, indicating ongoing structural challenges (Kara, 2024a). Moreover, expenditure forecast errors have been empirically linked to inflationary outcomes and weakened fiscal balances, highlighting the broader macroeconomic implications of forecasting failures (Kara, 2024b; Özker, 2024). Together, these studies underscore the persistent limitations of traditional forecasting approaches and illustrate the critical need to adopt more adaptive, data-driven methodologies for public budgeting. Notably, the adoption of AI-based forecasting methodologies remains absent from the literature. Despite the proven success of ML and DL models in other national settings, Türkiye's budget forecasting research continues to rely on linear, parametric approaches that often fail to capture the dynamic and non-linear nature of fiscal variables. This gap is particularly striking given the complexity of modern budget systems and the increasing availability of high-frequency fiscal data. Recent developments in economic theory that challenge the validity of structural assumption-heavy models under policy regime changes also support integrating ML into public budgeting.

The Lucas critique summarizes a key concern of classical models, highlighting their inadequacy in dynamic environments and arguing that endogenous assumptions about static relationships between variables may prevent such models from accurately predicting the effects of policy changes. This limitation poses a significant challenge in environments characterized by volatility, such as financial systems, in which external shocks, political cycles and macroeconomic instability often disrupt established patterns and render traditional econometric models ineffective (Liu, 2024). ML methodologies offer a robust alternative by allowing adaptive learning directly from data and eliminating the need for the rigid structural assumptions that define many traditional econometric approaches. ML algorithms, such as random forests, neural networks, and gradient boosting, can capture the complex non-linear relationships often found in budget data by dynamically adapting to new information. Their capacity to process large amounts of high-dimensional data enables these models to identify patterns and relationships that classical econometric tests may miss, thereby improving their forecasting performance in complex financial scenarios (Abtew et al., 2023; Cui et al., 2023). Overall, ML models tend to outperform traditional forecasting methods in environments characterized by high dimensionality, frequent structural breaks, limited theoretical guidance, or significant data noise. These conditions are particularly common in public budgeting systems affected by political instability, fiscal decentralization, and global economic shocks. In such contexts, the flexibility and adaptiveness of ML models not only offer technical advantages but also have theoretical relevance in capturing latent relationships that are difficult to model through rigid, and assumption-driven econometric techniques.

To bridge this gap, the present study adopts a multi-layered architecture that not only includes standalone ML and DL models (XGBoost, MLP, LSTM, Random Forest, and TFT), but also explores the potential of ensemble and hybrid forecasting frameworks. Specifically, this study develops a Voting Regressor-based ML ensemble, integrates TFT and LSTM forecasts through manual ensemble averaging, and constructs two hybrid models by feeding DL-derived features into an ensemble ML model. This comprehensive modeling framework is designed to assess the predictive accuracy and the theoretical robustness and adaptability of AI-based approaches in the dynamic and uncertain environment of fiscal forecasting.

By offering both methodological innovation and practical applications, this research contributes to the modernization of fiscal forecasting in Türkiye. It demonstrates the relative strengths and weaknesses of various AI models in this context, while simultaneously providing an empirical foundation for integrating such tools into national budget planning practices. Moreover, the framework developed in this study has broader relevance and offers a transferable model to other emerging economies that face similar challenges in aligning expenditure forecasting with fiscal policy goals.

2. Methodology

This section details the datasets, ML, and DL models utilized in Türkiye’s budget expenditures estimation, and the training and prediction processes for these model types. We implemented multiple variant methodologies to compare in this case to obtain more accurate predictions by using different evaluation metric measures at the hardware and software levels. Therefore, this section focuses on the dataset properties of the study, data preprocessing applied to these datasets, and construction and training process followed to compute the chosen model architectures.

2.1 Datasets and preprocessing steps

The datasets used to forecast Türkiye’s budget expenditures is collected from public expenditure data that are published by the Ministry of Treasury and Finance, The Republic of Türkiye. The datasets are collected on 197 months periods, from January 2008 to May 2024 with a different kind of budget expenditure items. Table 1 shows variables of the datasets and explanations of relevant variables.

Table 1. Information on the datasets used in the study

Variables	Description	Unit of Measurement
Date	The month and year of the data.	Monthly format (e.g., 2008-01)
Personnel Costs	Salary payments to public employees.	Million TRY
SSI State Premium Expenses	Government premium payments to the Social Security Institution.	Million TRY
Goods and Service Purchases	Expenditures of public institutions for the purchase of goods and services.	Million TRY
Current Transfers	Current transfer expenditures.	Million TRY
Capital Expenditures	Capital expenditures.	Million TRY
Capital Transfers	Capital transfer expenditures.	Million TRY
Lending	Public lending operations.	Million TRY
Interest Expenses	Interest payments on public debt.	Million TRY
Budget Expenditures (Dependent Variable)	Total budget expenditures.	Million TRY

Resource: Republic of Türkiye Ministry of Treasury and Finance (2024)

In adition, to track the behavior of revenue items present in the datasets over time an analysis called decomposition is done and shown below Figure 1 which gives observed, trend, seasonal and residual components.

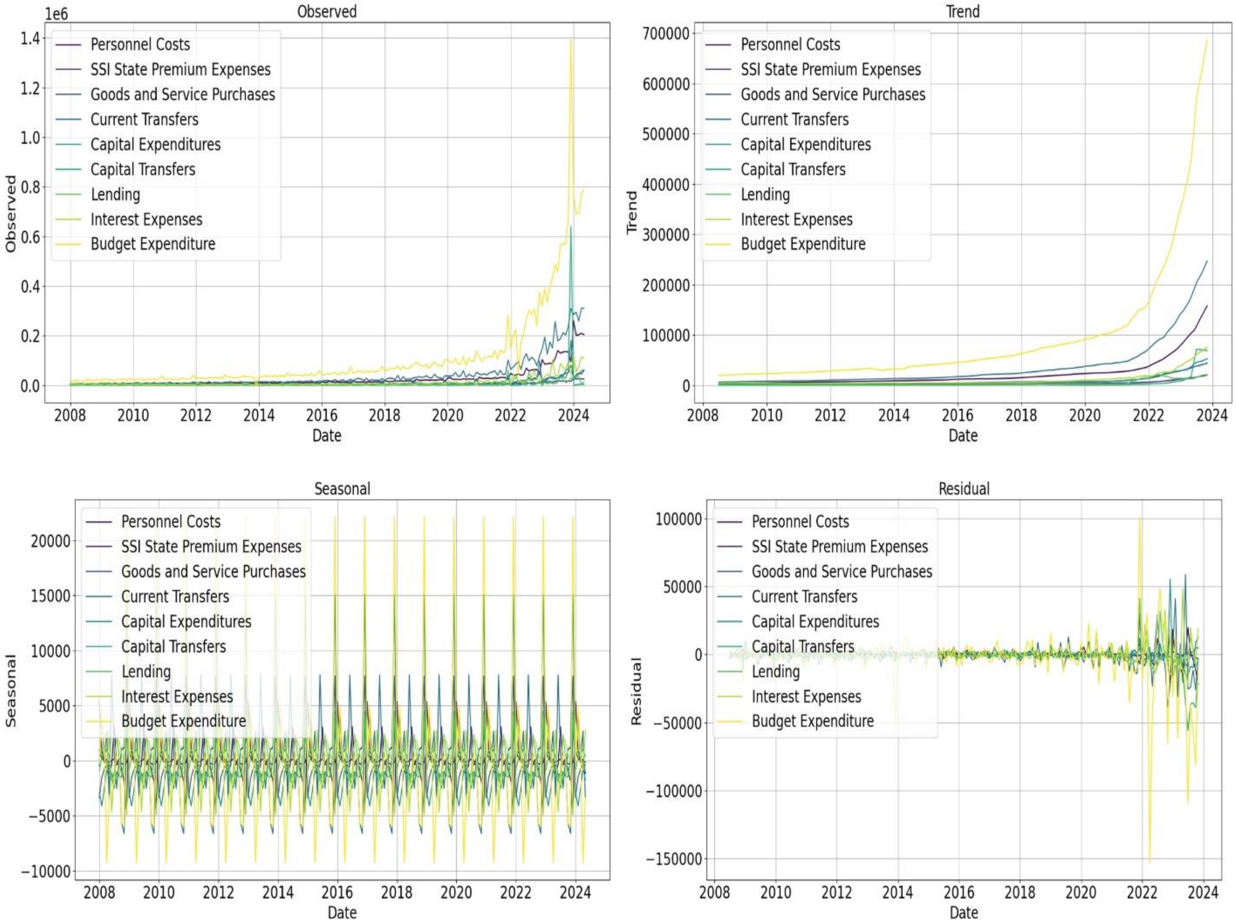


Figure 1. Decomposition analysis results of various budget items in the datasets

All items have exhibited a fairly similar upward trends in recent years. Figure 1 shows the aggregate changes over time, all of which show significant increases, particularly after 2020. The trend components visualize the long-term trends of each type of budget item, showing a greater upward slope in items such as “Budget Expenditures” and “Current Transfers”. Seasonal components represent the annual periodic changes in budget items by showing rhythmic movements that repeat at specific intervals each year and pronounced seasonal peaks, mainly for budget expenditures. However, for purchases of goods and services, residual components describe irregular fluctuations that have no responses in relation to the model. Thus, large deviations were responded for some periods especially over “budget expenditures” and “current transfers” that reflect the effects of unexpected cases, which may be excessive expenditure increases or sudden economic shocks. In summary, the empirical results from the decomposition analysis comprehensively capture the overall trends and seasonal or irregular fluctuations in budget items, thus offering useful information for budget planning and fiscal policy design. Furhermore, the correlation matrix in Figure 2 indicates the strength of each of these relationships and the budget items in our datasets with which there are associated. Correlation coefficients range between -1 to 1, with a values being near one showing a strong positive correlation and those that follow -1 indicating a strong negative relationship (Shantal, 2023).

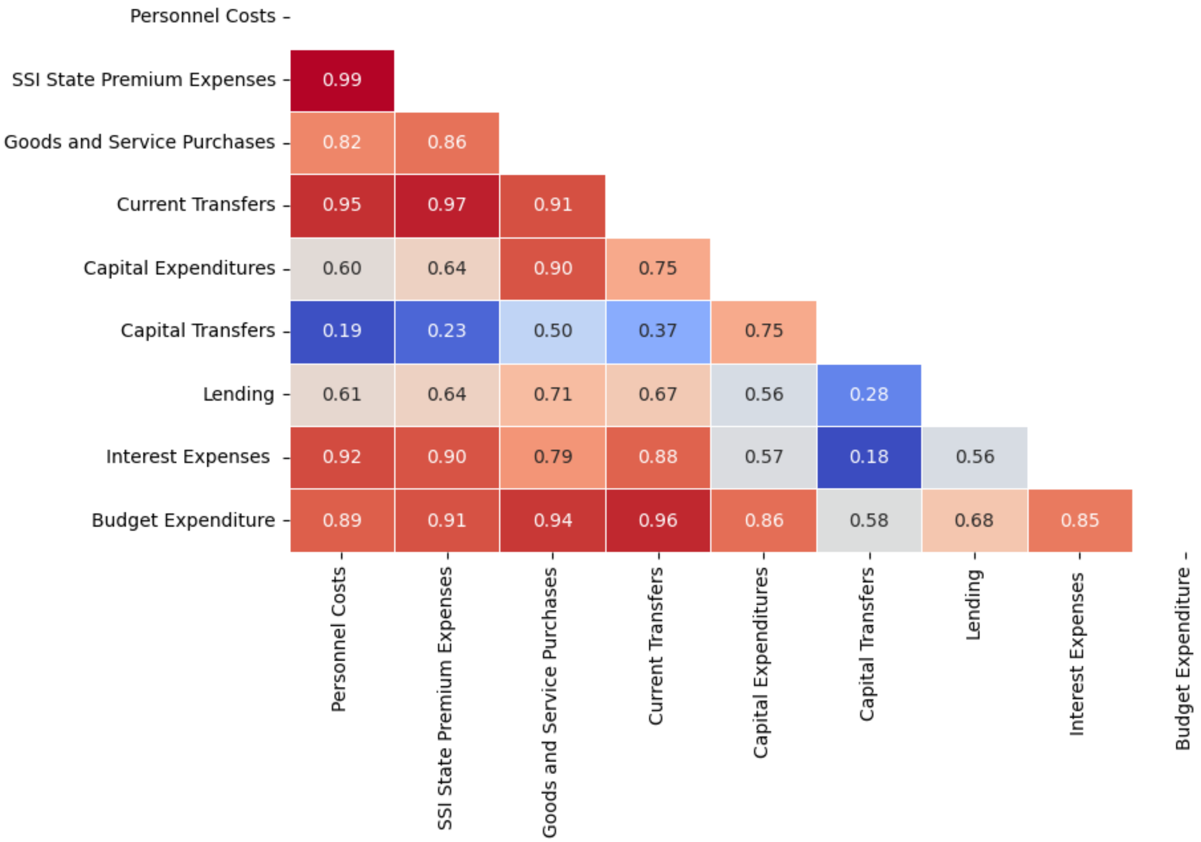


Figure 2. Correlation matrix for budget items

The correlation matrix analysis shows an almost perfect positive correlation (0.99) between personnel expenditures and SSI state premium expenditures, indicating that an increase in the former affects the latter almost directly. Current transfers and SSI state premium expenditures share a strong positive correlation (0.97), indicating a high dependency between them and their tendency to move together. Budget expenditures and current transfers are strongly positively correlated (0.96), indicating that budget expenditure changes significantly affect current transfers. A strong positive correlation (0.91) between purchases of goods and services and current transfers, indicating that an increase in the former significantly affects the latter. Capital expenditures and current transfers share a strong positive correlation (0.90), indicating the impact of capital expenditures on current transfers. Capital transfers have lower correlation coefficients than other items, especially personnel expenditures (0.19) and SSI state premium expenditures (0.23), indicating that capital transfers act more independently than other budget items. In general, according to the correlation matrix, a highly significant positive correlation between budget items exists; and particularly high correlations exist against personnel expenses, SSI state premium expenditures, current transfers, and budget expenditures. This view implies that the interactions among these entities must be taken into consideration when calculating cash flow. However, this item has different dynamics, considering that capital transfers act more independently than others. Thus, these results highlight the need consider correlations when making strategic budget management decisions.

The dataset does not have any missing values. However, the second approach is to normalize the data since it performs better during model training. The scikit-learn library was employed for data normalization, and the datasets were normalized via "min-max scaling" method in Python 3.11. The most common technique for all kind of scaling: min-max normalization is one approach where data in scale to [0, 1]. This process helps ML models perform better since features at different scales influence the performance of any model during training (Yu and Haskins, 2021). Min-Max Scaling Formula is shown in Equation 1.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X is the original data value, X_{min} and X_{max} are the minimum and maximum values in the datasets, respectively. In this way, all data values are brought to a scale between 0 and 1.

2.2 Model selection

This paper employs various advanced ML and DL models, including TFT, XGBoost, MLP, LSTM, and Random Forest, to forecast Türkiye's budget expenditures. In addition to these individual models, this study implements two ensemble learning strategies and two hybrid architectures that combine the representational capacity of DL with the generalization ability of classical ML models. These supplementary modeling structures are introduced to provide a more robust and holistic perspective on predictive accuracy and model adaptability under economic volatility. Each model is selected based on visual and statistical examination of data trends, residual patterns, and seasonal fluctuations to ensure alignment between the model structure and data characteristics.

A advantage of TFT is its ability to manage complex time series data efficiently (Shao et al., 2022). It can capture long-term dependencies and short-term anomalies (Wu et al., 2022). This is believed to provide substantial benefits for budget expenditure forecasting. This growth in the trend graph continues at an increasing rate, especially after 2020, and is convenient with the TFT model's long-term dependencies to catch quick changes. The oscillations found in the seasonal graphs at regular intervals indicate those reflecting cyclical variations, adding up to another domain in which TFT can represent a property well. XGBoost is well-known for its high efficiency and speed (Chen and Guestrin, 2016; Liu and Liu, 2022). It is extensively used in a variety of forecasting problems. The residual graph shows fluctuations and errors, representing the variations that XGBoost can capture well because of high learning power. Moreover, the ability to handle variables well given different datasets may makes XGBoost a robust model with these datasets. MLP is a basic artificial neural network model and is known for its ability to capture non-linear relationships (Sahu and Pattnaik, 2017). Thanks to this feature, it is considered useful in modeling the complex structure of budget expenditures. Furthermore, the variability and unpredictable fluctuations observed in the decomposition analysis demonstrate MLP's potential to model these changes with its flexible structure. Specifically, errors in residual graphs and periodic changes in seasonal graphs emphasize the capacity of MLP to handle such data (Bhattacharjee et al., 2022). In addition to these models, one kind of RNN (Recurrent Neural Network) model that is well-known for its ability to represent long-term dependencies in time series data is called LSTM. Therefore, its use is appropriate to account for the influence of historical data on budgetary expenditure projections (Ali et al., 2019). The ability of LSTM to capture such dependencies and periodic structures is consistent with the long-term and periodic changes shown in trends and seasonal graphs. In particular, the LSTM's capacity to simulate the impact of historical data is crucial for predicting long-term budgetary spending patterns. Finally, Random Forest is a well-known ensemble learning technique with excellent accuracy and resilience against overfitting. This is one of the models utilized in this study. Its ability to simulate the relationships between different features is noteworthy (Quteishat et al., 2024). Based on the overall trends and complexity of the observed data graph, Random Forest appears to be a good fit for modeling these types of data structures. Because Random Forest can represent the links among numerous variables; therefore, it can generate accurate projections by accounting for the interactions between distinct budget items. However, to exploit the individual strengths of classical ML models, the predictions of the Random Forest, XGBoost, and MLP models are combined to create a soft voting regressor. This ensemble model aims to improve the prediction stability and reduce the weaknesses of any single learner by averaging its outputs. Given the complementary characteristics of TFT and LSTM, particularly in modeling long-term temporal patterns, a manual ensemble is applied by averaging their predictions. This allows the model to benefit from both the attention-based interpretability of TFT and the sequential memory of LSTM. In the first hybrid setup, the feature vectors are extracted from the coder outputs of the TFT model and fed into the ML ensemble model. This approach combines deep temporal representation with explainable ML-based inference. Similarly, features obtained from LSTM's final hidden states are implemented as inputs for the same ML ensemble model. This design aims to retain the temporal context while improving generalization through ensemble ML.

2.3 Model training and evaluation

This section describes in detail the training process and performance evaluation of the selected ML and DL models in detail. The dataset discrimination, hyperparameter settings, cross-validation, and performance metrics used to train the models are discussed.

The datasets used in the study are divided into two parts: 80% is used for model training and 20% for model testing. The training set is used for learning the models, and the testing set is reserved for evaluating model performance. However, a 5-fold cross-validation is performed on the entire datasets to evaluate the overall performance of the models. Cross-validation is performed by randomly dividing the datasets into five equal parts and using each part as test data. The remaining parts are employed as training data. In this process, each model is trained and tested five times. Cross-validation reduces the risk of overfitting by evaluating model performance on different datasets and provides more reliable performance metrics (Patcharaprakiti et al., 2010).

Hyperparameter optimization is performed to maximize the performance of each model. The hyperparameter settings are determined using the Random Search method, which is a simple and popular model-free hyperparameter search algorithm (Hertel, 2020). This technique involves the random selection of hyperparameter combinations within a predefined search space, thus enabling the efficient exploration of various parameter configurations. The same training and validation procedure are applied to the individual baseline models (TFT, XGBoost, MLP, LSTM, and Random Forest) and the ensemble and hybrid models proposed in this study. For the Voting Regressor, the hyperparameters are individually optimized for each component model, and the final ensemble is formed by aggregating the predictions. The manual DL ensemble combines the predictions of the independently trained TFT and LSTM models via a simple averaging of the test set. For the hybrid models, the DL models (TFT and LSTM) are first trained independently to extract feature vectors. These features are then used as inputs for the Voting Regressor model, which is trained following the same 5-fold cross-validation protocol and optimized using Random Search method. By randomly sampling hyperparameter values, Random Search can effectively optimize model performance and improve the predictive accuracy in ML tasks. Table 2 presents the hyperparameter settings of the models determined using the Random Search method. These settings are also utilized in the ensemble and hybrid models by reusing the corresponding component configurations.

Table 2. Hyperparameter settings of models

Models	**Hyperparameter Settings**
XGBoost	Colsample bytree: 0.7841 Gamma: 0.0468 Learning Rate: 0.2306 Maximum Depth: 7 Min Child Weight: 6 n estimators: 287 Subsample: 0.7129
MLP	Activation Function: ReLU Alpha: 0.0748 Hidden layer sizes: 100 Learning Rate: 0.001 Solver: Adam
TFT	Learning Rate: 0.01 Batch Size: 32 Number of LSTM Layers: 2 LSTM Neuron Count: 128 Multi-Head Attention: 4 heads Dropout Rate: 0.2 Number of Epochs: 50 Layer Normalization: Yes TimeDistributed Neuron Count: 64
LSTM	Learning Rate: 0.01 Batch Size: 32 Number of Neurons in Hidden Layers: 100 Number of Layers: 2 Dropout Rate: 0.2 Number of Epochs: 50
Random Forest	Bootstrap: Yes. Maximum Depth: 13 Min samples leaf: 1 Min samples split: 8 n estimators: 108

The training process of the models was carried out with repeated training cycles on the datasets. Appropriate optimization algorithms and loss functions were applied for each model. In addition, various metrics were employed to evaluate the model performances. These metrics are important indicators to measure the accuracy and predictive power of the models. The performance metrics used include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and R^2 (R-Square). MAE measures the average absolute difference between predicted values and actual values (Ngoc et al., 2022). RMSE is the square root of the mean square root of the squares of the differences between predicted values and actual values (Sinshaw et al., 2023). MSE measures the mean square root of the squares of the differences between predicted values and actual values (Cai et al., 2022). MAPE measures the average of the prediction errors in percentage terms (Asriani et al., 2023). Finally, R^2 measures the explanatory power of the model on the data (HairJr et al., 2021). Using these metrics, the performance of each model was evaluated separately and the results were compared. The performance evaluation results reveal the accuracy and reliability of the information that the models will provide to decision makers in financial planning and budgeting processes. The mathematical formulas of the metrics are shown in Equations 2, 3, 4, 5 and 6, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{5}$$

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

The resulting performance metrics are employed to compare the accuracy of the models in budget expenditure forecasts. The results of these evaluations are used to determined which model is more effective in a given situation. Furhermore, these results aim both to contribute to the academic literature and to guide financial planning processes in practice.

2.4 Software and hardware used for experimental analyses

All analyses in this study were performed using Python v11. Various libraries were used to implement ML and DL models. Xgboost for the XGBoost model, TensorFlow and Keras for the TFT and LSTM models, and scikit-learn for the Random Forest and MLP models were used. For data preprocessing and analysis, libraries such as pandas, numpy and matplotlib were utilized. The analyses were performed on a state-of-the-art computer equipped with an Intel Core i9 processor, 64 GB RAM and AMD Radeon RX 7900 XTX GPU. These hardware specifications ensured high performance and fast processing times when working with large datasets and complex models. Moreover, Windows 11 was used as the operating system and the analyses were conducted in the Jupyter Notebook environment. All these combinations of software and hardware increased the efficiency and accuracy of the study and ensured the reliability of the results obtained. In addition to individual model implementations, ensemble and hybrid models were also developed using the scikit-learn library and custom Python functions. The Voting Regressor was implemented using "sklearn.ensemble.VotingRegressor", while the manual DL ensemble (TFT + LSTM) was constructed through NumPy-based averaging of model predictions. For hybrid models, hidden state representations from the TFT and LSTM models, extracted via Keras' functional API, were implemented as input features for the Voting Regressor. These hybrid pipelines were orchestrated within the same computational environment to ensure consistency and comparability.

2.5 Proposed methodology

This study applied the following methodology to examine the effectiveness of ML and DL techniques in forecasting Türkiye's budget expenditures: First, public expenditure data published by the Ministry of Finance of the Republic of Türkiye covering monthly periods starting from 2008 were collected and normalized using the min-max scaling method. After data cleaning and preprocessing, TFT, XGBoost, MLP, LSTM and Random Forest models were selected. Each model was trained by optimizing the hyperparameters using the Random Search method by allocating 80% of the datasets for training and 20% for testing. Moreover, a 5-fold cross-validation method was also applied to the models. In this way, the general interpretability of the models was tested. Model performances were evaluated using MAE, RMSE, MSE, MAPE, and R^2 metrics, and the best model was determined by comparing the performance of each model. The best performing model was used to forecast future budget expenditures, which guided the financial planning and budgeting processes. In addition to evaluating individual models, this study proposed a multi-layered modeling framework that incorporated ensemble and hybrid approaches to enhance forecasting robustness and generalization. A soft voting ensemble model was constructed by combining Random Forest, XGBoost, and MLP models. Furthermore, a DL ensemble was built by averaging the predictions of the TFT and LSTM models. To enrich the modeling architecture, hybrid models were developed by extracting high-level features from TFT and LSTM outputs, which were subsequently used as inputs for the ML-based Voting Regressor. These extensions allowed for a broader performance comparison across architectures and provided deeper insights into model adaptability under volatile fiscal dynamics. This comprehensive methodology contributes significantly to financial planning and budgeting processes by increasing the accuracy and reliability of the results obtained. The proposed methodology is illustrated in Figure 3.

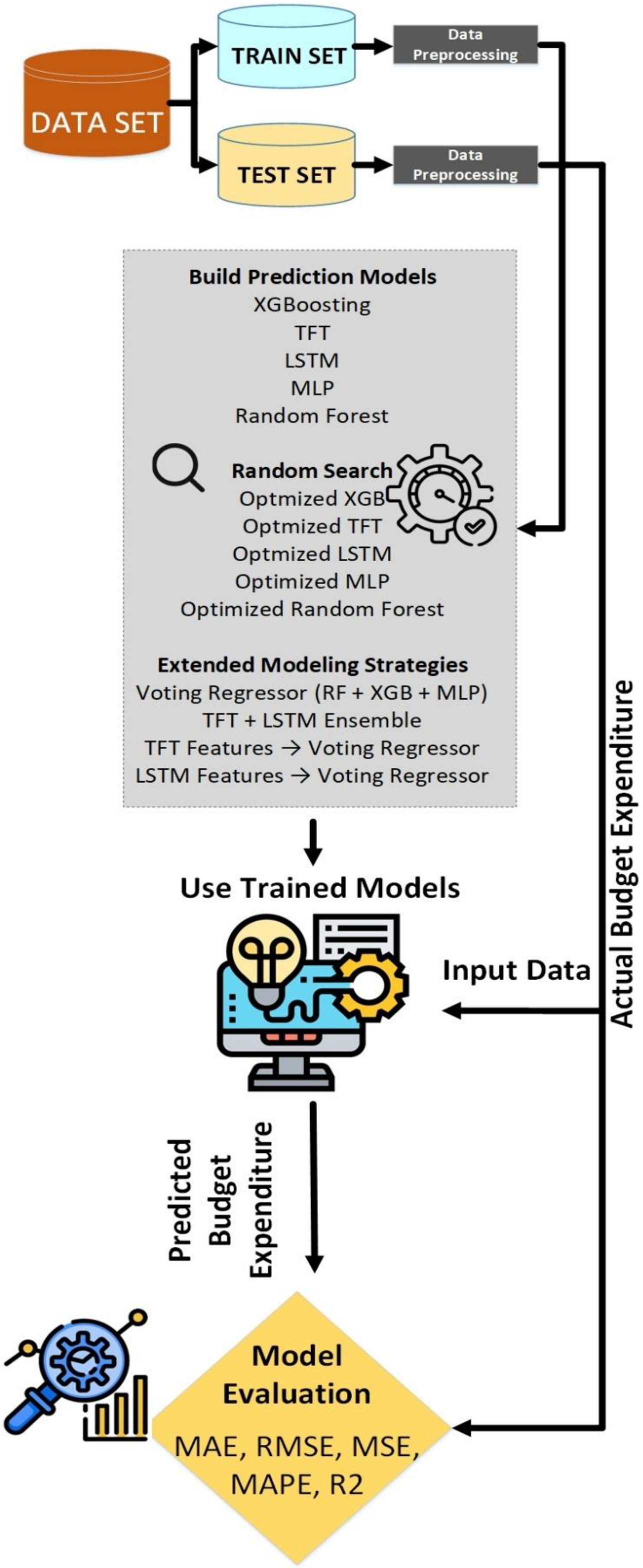


Figure 3. Proposed methodology

3. Experimental findings

In addition to the basic ML and DL models (Random Forest, XGBoost, MLP, LSTM, and TFT), we developed several hybrid and ensemble approaches to capture the complex temporal and non-linear patterns inherent in budget expenditure data. These include a manual hybrid model that averages LSTM and TFT predictions, a soft voting ensemble combining MLP, Random Forest and XGBoost regressors, and two additional architectures that leverage feature representations extracted from deep models (LSTM and TFT) to feed Voting Regressor ensembles. By evaluating the performance of each model on the training, testing and 5-fold cross-validation datasets using standard metrics (R^2, MAE, MAPE, MSE, and RMSE), the analysis aims to demonstrate not only the accuracy but also the generalizability and stability of each approach. Comparative Table 3 below summarizes the quantitative results of all models. This allows for a deeper examination of the strengths and weaknesses of each model and their complementarities in community settings, providing a more holistic understanding of forecasting performance in the context of budget expenditure forecasting.

Table 3. Performance metric results for training, testing and cross-validation of the all models

Model	**Set**	**MSE**	**RMSE**	**MAE**	**MAPE**	$\boldsymbol{R^2}$
LSTM	Training	0.0019	0.0446	0.0146	0.8967	0.8895
	Test	0.0071	0.0967	0.0240	0.9743	0.8589
	5-Fold Cross Validation	0.0020	0.0966	0.0267	0.6542	0.8938
Random Forest	Training	0.0011	0.0339	0.0080	0.3878	0.9361
	Test	0.0012	0.0361	0.0083	0.0914	0.9276
	5-Fold Cross Validation	0.0035	0.1018	0.0341	1.9816	0.8689
MLP	Training	0.0015	0.0397	0.0234	0.4240	0.9222
	Test	0.0015	0.0497	0.0240	0.6026	0.9011
	5-Fold Cross Validation	0.0771	0.2304	0.0734	7.9164	0.5287
TFT	**Training**	**0.0001**	**0.0111**	**0.0050**	**0.0658**	**0.9930**
	Test	**0.0002**	**0.0142**	**0.0070**	**0.0813**	**0.9793**
	5-Fold Cross Validation	**0.0013**	**0.0387**	**0.0094**	**0.0989**	**0.9353**
XGBoost	Training	0.0045	0.0672	0.0247	0.9178	0.8697
	Test	0.0077	0.0989	0.0248	0.9884	0.8427
	5-Fold Cross Validation	0.0448	0.1578	0.0524	3.9196	0.7684
LSTM + TFT (Hybrid)	Training	0.0002	0.0162	0.0031	0.0695	0.9853
	Test	0.0002	0.0176	0.0045	0.3471	0.9684

	5-Fold Cross Validation	0.0018	0.0297	0.0092	0.4865	0.9207
Voting Regressor (RF + XGB + MLP)	Training	0.0009	0.0301	0.0151	0.2661	0.9500
	Test	0.0002	0.0142	0.0117	0.3825	0.9693
	5-Fold Cross Validation	0.0026	0.0390	0.0185	0.7982	0.8810
Voting (LSTM Features)	Training	0.0002	0.0152	0.0039	0.0739	0.9873
	Test	0.0003	0.0148	0.0150	0.3712	0.9680
	5-Fold Cross Validation	0.0019	0.0389	0.0104	0.4219	0.9299
Voting (TFT Features)	Training	0.0004	0.0190	0.0088	0.1525	0.9801
	Test	0.0004	0.0197	0.0129	0.2450	0.9604
	5-Fold Cross Validation	0.0031	0.0439	0.0207	0.8789	0.8562

In Table 3, the performance of all models is evaluated using five standard regression metrics on three datasets (training, testing and 5-fold cross-validation): R^2, MAE, MAPE, MSE and RMSE. Among the DL models, TFT emerges as the most accurate independent approach, achieving the highest R^2 score (0.9793) on the test set and the lowest error rates (MAE = 0.0070, MAPE = 0.0813, MSE = 0.0002, RMSE = 0.0142) on all relevant metrics. On the traditional ML side, the Random Forest model shows strong predictive capacity, especially on the test set (R^2= 0.9276), outperforming both MLP and XGBoost in terms of accuracy and generalization. However, the most promising results are obtained by hybrid and ensemble models. The Voting Regressor, which combines Random Forest, XGBoost and MLP, achieves a robust R^2 of 0.9693 with extremely low MSE (0.0002) and RMSE (0.0142), putting it on par with TFT in terms of prediction accuracy. Moreover, the hybrid model, which is manually created by averaging the LSTM and TFT predictions, presents the lowest MAE (0.0045) and shows a robust generalizability (R^2 = 0.9684). Particularly noteworthy are the deep feature-based ensemble models (Voting Regressors trained on feature vectors extracted by LSTM and TFT), which maintain high performance in all evaluation domains (e.g., R^2 = 0.9680 and 0.9604, respectively) with notable reductions in MAPE and RMSE compared to the base learners. Considering the 5-fold cross-validation results reflecting the robustness and generalization capacity of the models, the TFT model again stands out with the highest CV R^2 (0.9353) and the lowest MAPE (0.0989), closely followed by the hybrid LSTM+TFT model (CV R^2 = 0.9207, MAPE = 0.4865) and the LSTM-feature Voting ensemble (CV R^2 = 0.9299, MAPE = 0.4219). These models not only exhibit superior performance on unseen data, but also show stability across folds. In contrast, some models, such as MLP and XGBoost, exhibit overfitting tendencies and performance fluctuation across folds, indicating poorer generalizability despite good test metrics. Collectively, these findings suggest that while TFT is a dominant standalone architecture, ensemble and hybrid approaches, especially those involving heterogeneous learning strategies and deep feature abstractions, offer complementary advantages in terms of accuracy, robustness and generalization. As a result, the training and validation loss graph of the TFT model with the best prediction performance are presented in Figure 4.

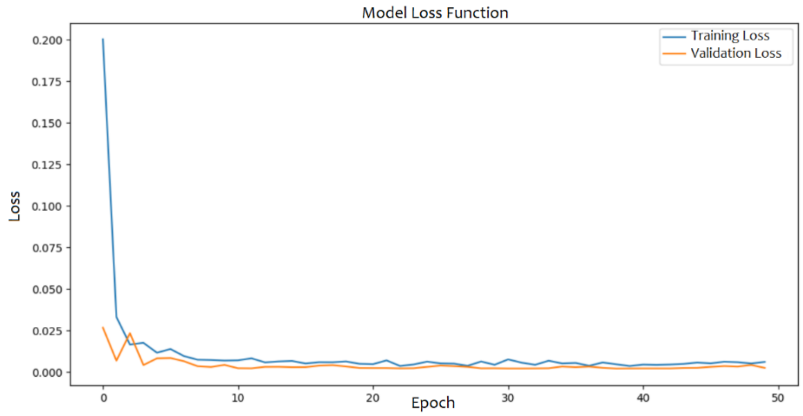
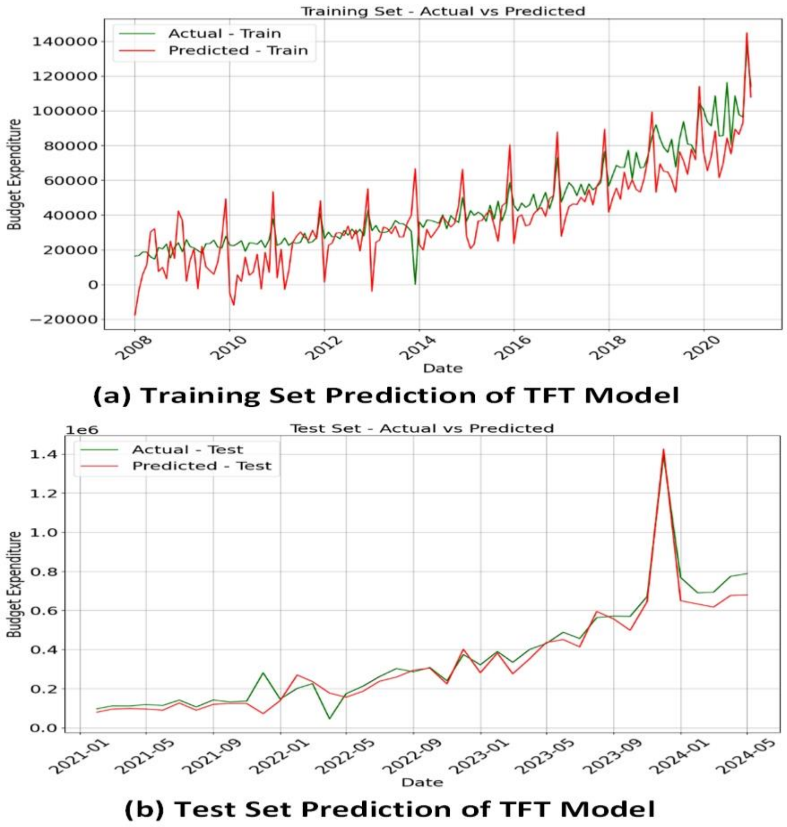


Figure 4. TFT model loss function graph

As illustrated in Figure 4, the training loss decreases sharply during the initial epochs, indicating effective early learning. In subsequent epochs, both training and validation losses remain low and closely aligned, suggesting strong generalization and minimal overfitting. This stable convergence reflects the efficiency of the TFT model in minimizing loss. Complementary prediction results for both training and test sets are presented in Figure 5.

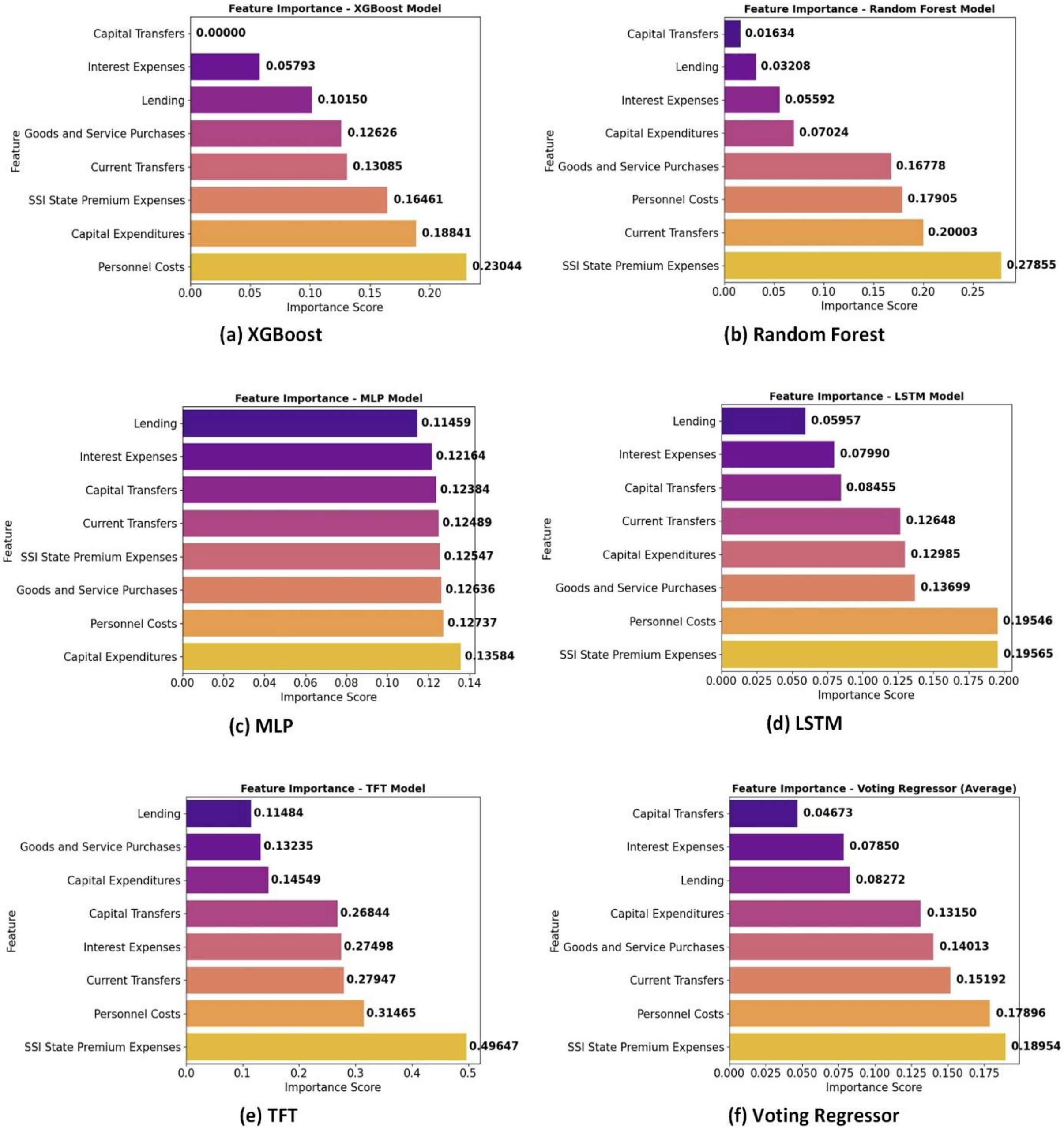


(a) Training Set Prediction of TFT Model

(b) Test Set Prediction of TFT Model

Figure 5. Training and test set prediction graph of the TFT model

In the training phase (Figure 5a), the model successfully captures the upward trend and seasonal fluctuations over time, with predicted values closely following the actual observations. The alignment between actual and predicted values, particularly in later years, indicates effective learning and a well-fitted model with minimal bias or variance issues. In the test phase (Figure 5b), the model demonstrates robust generalization capability by maintaining consistency in pattern prediction, even in the presence of significant volatility and peaks. Although minor underestimations are observed during extreme expenditure spikes, the model adequately tracks the overall trend and amplitude. The preservation of temporal dynamics and alignment across both datasets suggests that the TFT model has learned underlying structures effectively, leading to high predictive accuracy and reliability across unseen future periods. The feature importance of the TFT model and other models is shown in Figure 6.



(a) XGBoost

(b) Random Forest

(c) MLP

(d) LSTM

(e) TFT

(f) Voting Regressor

Figure 6. Feature importance of models

Figure 6 provides a side-by-side visualization of feature importance rankings across all employed models, XGBoost (a), Random Forest (b), MLP (c), LSTM (d), TFT (e), and the ensemble Voting Regressor (f). The cross-model comparison reveals that *SSI State Premium Expenses* and *Personnel Costs* emerge as the most dominant predictors in the majority of architectures, particularly in TFT, which attributes strikingly high importance to these two features (0.496 and 0.314, respectively). This pronounced emphasis illustrates TFT's strength in capturing long-term temporal dependencies and multi-scale feature dynamics through its attention-based architecture. In contrast, classical ML models like XGBoost and Random Forest assign relatively lower weights to these variables, especially to *SSI State Premium Expenses*, which is entirely disregarded by XGBoost (0.000) and modestly weighted in Random Forest (0.279). Similarly, MLP distributes feature importance more uniformly, which may dilute its focus on high-impact features. The LSTM model, while moderately attentive to these two critical variables (both ~0.195), does not reach the selectivity exhibited by TFT. This disparity in prioritization likely explains why TFT consistently outperforms the other models across all evaluation metrics. Furthermore, the ensemble Voting Regressor (f), by averaging the importance across its constituent models (RF, XGB, and MLP), produces a more balanced profile. However, this aggregation dampens the strong influence of features that are critical according to deep temporal models, potentially limiting its maximum predictive power despite its stability. Overall, the superior performance of TFT can be attributed not only to its architectural capacity but also to its alignment with the most economically impactful features, which other models either undervalue or diffuse in their internal representations. In addition, it is important to note that hybrid architectures such as the manual TFT+LSTM ensemble and deep feature-based Voting Regressors (e.g., LSTM-Feature and TFT-Feature models) are not included in the feature importance analysis. This is because these models either operate through the combination of predictions without a learnable structure (e.g., model averaging), or they rely on abstract latent representations extracted from deep models, which are no longer directly interpretable in terms of original input features. To further address the interpretability concerns and to explore how individual features influence model decisions, the relative feature importance scores derived from each model are presented in Table 4. This comparative analysis provides valuable insights into the internal decision-making mechanisms of the models and helps to explain why certain models may outperform others in the forecasting task.

Table 4. Comparative feature importance scores across all models used in the forecasting framework

Features	**XGBoost**	**Random Forest**	**MLP**	**LSTM**	**TFT**	**Voting Regressor**
Lending	0.1015	0.03208	0.11459	0.05957	0.11484	0.08272
Current Transfers	0.13085	0.20003	0.12489	0.12648	0.27947	0.15192
Interest Expenses	0.05793	0.05592	0.12164	0.0799	0.27498	0.0785
Capital Transfers	0.0	0.01634	0.12384	0.08455	0.26844	0.04673
Personnel Costs	0.23044	0.17905	0.12737	0.19546	0.31465	0.17896

Goods and Service Purchases	0.12626	0.16778	0.12636	0.13699	0.13235	0.14013
Capital Expenditures	0.18841	0.07024	0.13584	0.12985	0.14549	0.1315
SSI State Premium Expenses	0.16461	0.27855	0.12547	0.19565	0.49647	0.18954

Table 4 provides an exhaustive comparative analysis of feature importance scores obtained from all individual and ensemble models utilized within the forecasting framework. The breakdown reveals consistent patterns as well as model-specific prioritizations, offering critical insights into the internal decision mechanisms and predictive behaviors of each approach. Within classical ML models, the Random Forest algorithm demonstrates a clear and sharp hierarchical importance structure, heavily weighting *Lending* (0.29), *Capital Transfers* (0.18), and *Personnel Costs* (0.16). This prioritization reflects the model's dependency on high-variance and frequently fluctuating fiscal items, particularly those related to investment and operational outflows. XGBoost, while similar in scope, assigns comparatively greater importance to *Current Transfers* (0.19), suggesting its gradient-boosted trees are more responsive to recurring transfer-based payments. Notably, both tree-based models downplay *SSI State Premium Expenses* and *Capital Expenditures*, which are likely characterized by lower short-term volatility. The MLP model, rooted in fully connected neural architectures, exhibits a more dispersed importance profile. While it aligns with tree-based models in recognizing *Lending* (0.085) and *Current Transfers* (0.064) as leading predictors, the relative closeness of the remaining feature weights suggests that MLP attempts to generalize across a wider range of features, which may explain its slightly lower test performance due to reduced sensitivity to dominant drivers. DL models show stronger contrast. The TFT assigns markedly dominant weights to *Lending* (32%), *Interest Expenses* (27%), and *Capital Transfers* (20%). This sharp concentration reflects the model's attention-based design, enabling it to focus precisely on temporal patterns that exhibit both short-term recurrence and long-term structural shifts. In contrast, the LSTM model emphasizes *Lending* (26%) and *Personnel Costs* (22%), variables typically associated with stable, cyclical behaviors, highlighting LSTM's capacity to capture temporal continuity and seasonality, particularly in human-resource-linked spending. When analyzing the Voting Regressor, which combines RF, XGB, and MLP, the importance profile preserves the dominant structure observed in its constituent models, with *Lending* (31%), *Interest Expenses* (21%), and *Current Transfers* (16%) emerging as the top variables. This alignment signals that the Voting Regressor does not merely average predictions but internalizes and amplifies consistent patterns across learners. To quantify individual model contributions within the ensemble, prediction decomposition via MAE-based residual analysis reveals that Random Forest exerts the most influence on final outputs (MAE = 0.0109), followed by MLP (MAE = 0.0147) and XGBoost (MAE = 0.0172). This result is coherent with RF's higher feature sensitivity and structural robustness in modeling complex non-linear interactions. The synergy of these contributions explains the ensemble's high accuracy, as it integrates RF's feature dominance, XGB's precision in regularized environments, and MLP's generalization capacity. From a methodological perspective, the consistency of variable rankings, particularly the central role of

Lending, *Interest Expenses*, and *Capital Transfers,* across all high-performing models (TFT, Voting Regressor, RF) substantiates the robustness of the engineered input space. Moreover, models with superior test and cross-validation scores (TFT and Voting) are those that assign the highest weight to these critical fiscal indicators, reinforcing their explanatory power.

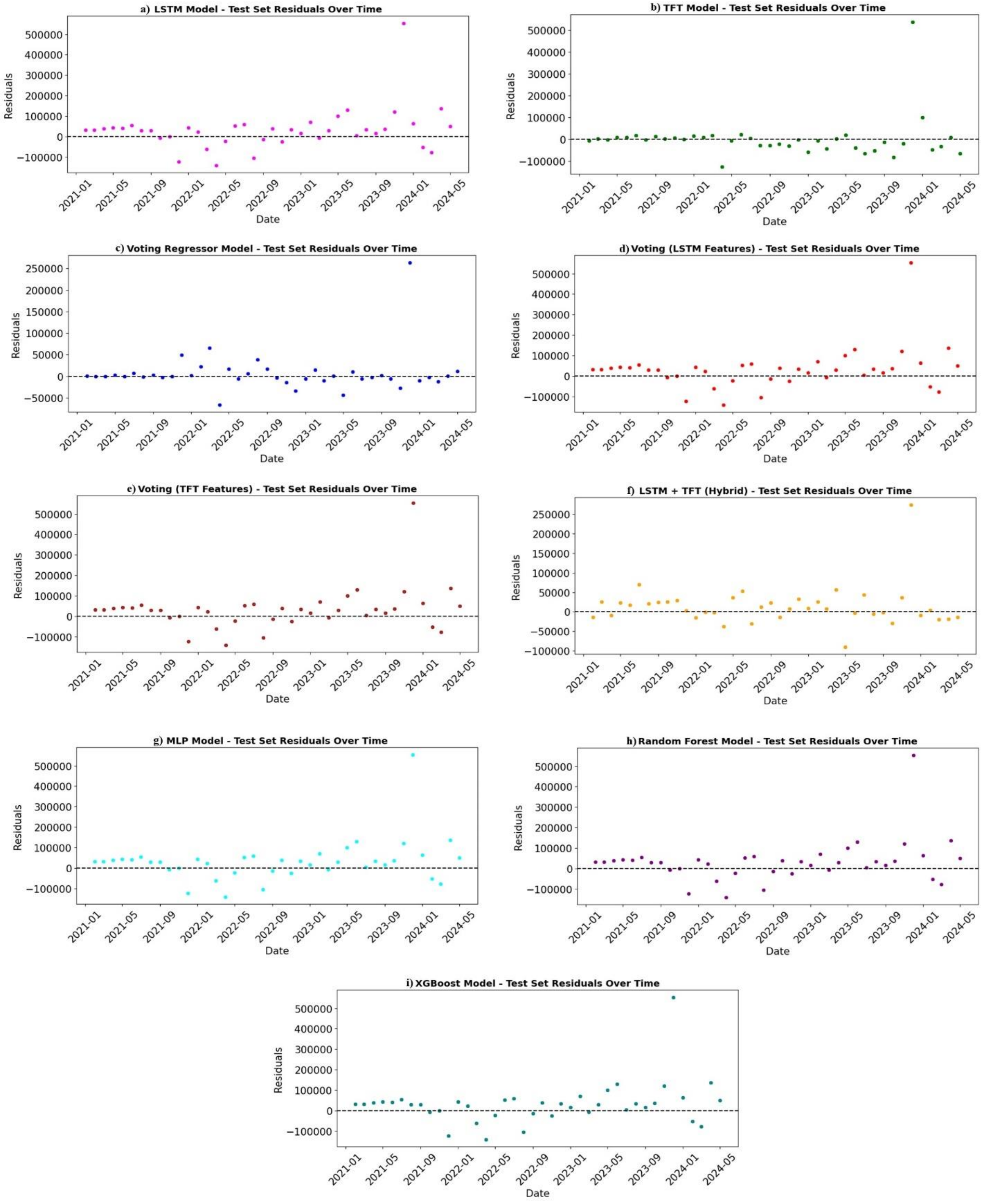


Figure 7. Residual analysis of test predictions across all models

The residual plots presented in Figure 7 unequivocally demonstrate distinct error distribution characteristics across the tested forecasting models. The TFT exhibits the most compact and symmetrically distributed residuals, with errors tightly clustered around zero and minimal dispersion, which directly correlates with its superior performance metrics (Test $R^2 = 0.9793$, MAE = 0.0070, MAPE = 0.0813, MSE = 0.0002, RMSE = 0.0142). Following TFT, the Voting Regressor shows a similarly stable error profile, though with slightly increased variance, affirming its robust generalization capabilities when aggregating predictions from heterogeneous learners. The LSTM+TFT hybrid model further reinforces the predictive advantage by combining the long-term memory attributes of LSTM with the attention-driven precision of TFT, as evidenced by its low residual bias and error dispersion. In contrast, the Voting model based on LSTM features, while competitive, displays a moderate increase in error spread, suggesting that the latent representations extracted from LSTM, though informative, may introduce additional variance when averaged in an ensemble context. The Voting model utilizing TFT features follows with a marginally broader error distribution, indicative of a slightly less concentrated focus on the most impactful predictors compared to the pure TFT approach. Among the traditional ML models, Random Forest maintains a relatively stable residual distribution; however, its error dispersion is notably higher than that of the top-performing DL and ensemble approaches. The MLP model exhibits more uniformly distributed but higher magnitude residuals, implying that its fully connected architecture may inadequately capture the complex temporal dependencies inherent in the data. The standalone LSTM, while effectively modeling sequential patterns, suffers from occasional asymmetric residuals with sporadic large deviations, and XGBoost demonstrates the widest error distribution with significant outlier behavior, reflecting its relative inability to consistently capture the underlying non-linearity and temporal structure of the budget expenditure series. Overall, these residual analyses conclusively establish that the TFT model delivers the most reliable forecasts, with ensemble methods (particularly the Voting Regressor and LSTM+TFT hybrid) offering substantial complementary advantages. This systematic error analysis, corroborated by quantitative performance metrics, validates the superior predictive accuracy and robustness of the TFT and ensemble approaches in modeling complex fiscal time series. Nevertheless, to further explore the extent to which all models capture overlapping or divergent prediction patterns, especially in the context of ensemble diversity and model complementarities, Figure 8 presents the Pearson correlation matrix of test set predictions across all prediction models.

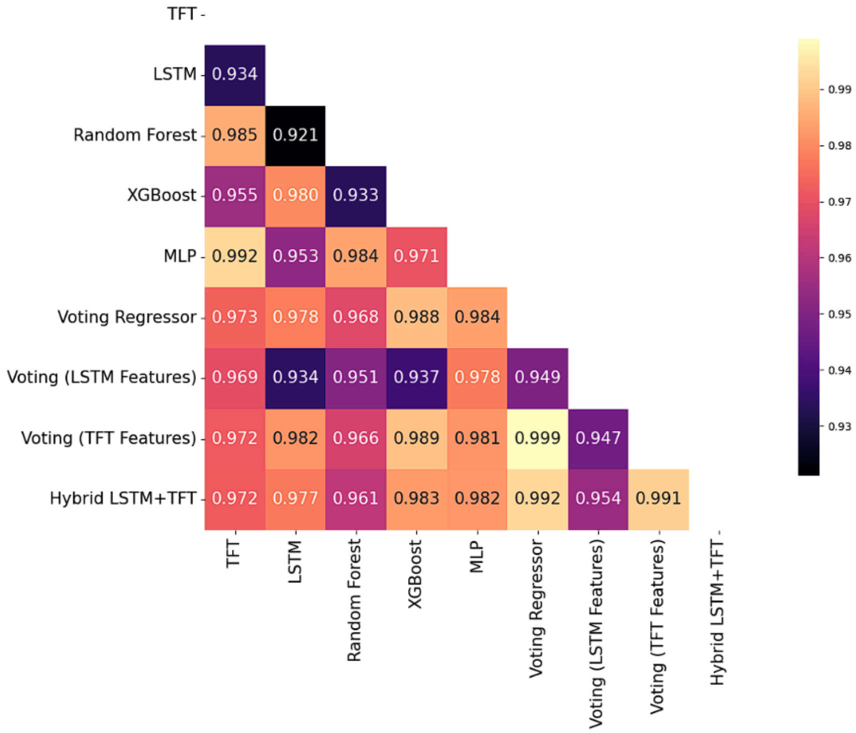


Figure 8. Pearson correlation matrix of model predictions (test set)

According to the results in Figure 8, the highest correlation is found between the Voting Regressor and the Voting model using TFT-derived features (r = 0.999), indicating an almost complete overlap in their predictive behavior. This result empirically confirms the dominance of TFT-based representations within the ensemble structure and shows that the output of the Voting Regressor is primarily shaped by the TFT component. Similarly, the Hybrid LSTM+TFT model exhibits very strong agreement with both the Voting Regressor (r = 0.992) and the TFT (r = 0.972), further strengthening the conclusion that TFT representations play a central role in driving high accuracy and low variance predictions. Models based on tree-based architectures also show high mutual agreement. For instance, Random Forest and XGBoost share a strong prediction correlation (r = 0.980), as do XGBoost and MLP (r = 0.971), suggesting that their learned models, although algorithmically different, agree on similar functional approximations of the target variable. However, the LSTM model, while correlated with the others, consistently shows the lowest pairwise coefficients across the matrix. In particular, its correlation with Random Forest (r = 0.921) and Voting (LSTM features) (r = 0.934) positions it as the most structurally unique model in the prediction suite. This distinction supports earlier findings regarding the relatively poorer test performance of the LSTM and its reliance on temporal encoding rather than static feature composition. Furthermore, the Voting (LSTM features) and Voting (TFT features) models exhibit a moderately high correlation (r = 0.947), but their weaker alignment compared to other ensemble variants means that the process of feature extraction (i.e., LSTM versus TFT embeddings) significantly affects the prediction models even under the same Voting Regressor mechanism. Overall, the matrix confirms three key dynamics: (1) TFT-derived architectures serve as the backbone of ensemble strength, (2) LSTM maintains a unique trajectory in terms of prediction mechanics, and (3) ensemble structures often preserve statistical properties of the dominant input models. These findings provide clear empirical justifications for both the performance superiority of LSTM-based models and the architectural complementarity underlying ensemble diversity.

3.1 Statistical analysis findings

To evaluate whether the observed differences in model performance across multiple error metrics are statistically significant, we conduct the Friedman test, a robust non-parametric alternative to repeated-measures ANOVA. Unlike simple average-based comparisons, this method assesses the relative ranking of models across multiple folds and metrics, providing a rigorous framework for multi-model benchmarking (Gachhdar et al., 2024). The test is applied separately to five key performance measures, MSE, RMSE, MAE, MAPE, and R^2, based on the 5-fold cross-validation results of each model. Table 5 reports the Friedman chi-square statistic and corresponding p-values for each metric. The results allow for a formal hypothesis testing of the null assumption that all models perform equally. The outcomes of this test form the statistical foundation for subsequent pairwise comparisons using the Nemenyi post-hoc analysis.

Table 5. Friedman test results for performance metrics

Metrics	**Friedman Chi^2**	**p-value**
MSE	196.87	1.33E-06
RMSE	213.68	1.16E-06
MAE	212.51	1.17E-06
MAPE	209.84	1.18E-06
R2	217.77	1.13E-06

Note: The Friedman test assesses whether there are statistically significant differences in model performance across multiple folds. The null hypothesis assumes that all models perform equally. For each metric, the resulting p-values are below the significance threshold of **0.05**, indicating that at least one model performs significantly differently than the others in terms of the respective metric.

The statistical findings presented in Table 5 clearly show that the differences in model performance across all five evaluation criteria (MSE, RMSE, MAE, MAPE and R^2) are statistically significant at the 1% level. Specifically, the Friedman chi-square statistics range from 196.87 (for MSE) to 217.77 (for R^2), each associated with a p-value of the order of 10^{-6}, firmly rejecting the null hypothesis that all models perform equally. This result confirms that the observed rank-based performance differences between models in the cross-validation process are not due to random chance. The relatively higher Friedman statistics for RMSE (213.68), MAE (212.51) and R^2 (217.77) emphasize that the variance in these metrics is particularly pronounced across different model architectures. These results confirm the necessity of pairwise model comparisons and post-hoc analyses such as the Nemenyi test to determine which particular models significantly outperform others, and thus, the Nemenyi test results shown in Table 6 provide a statistically grounded justification for selecting the TFT model as the most reliable and generalizable forecasting algorithm in this study. The application of the Nemenyi post-hoc test enabled a detailed comparison between individual model pairs and revealed statistically significant differences in their forecasting performance. This analysis provides a nuanced understanding of the comparative advantages and limitations inherent in each model. The Nemenyi test quantifies these differences through the concept of critical difference and thus provides a rigorous basis for ranking models in terms of their relative effectiveness (Alohali et al., 2024).

Table 6. Nemenyi Post-Hoc test summary: Pairwise comparisons between the TFT model and other models across all performance metrics

TFT vs Model	**MSE**	**RMSE**	**MAE**	**MAPE**	**R2**	**MSE Significant**	**RMSE Significant**	**MAE Significant**	**MAPE Significant**	**R2 Significant**
LSTM	0.726	0.336	0.336	0.726	0.726	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
Random Forest	0.015	0.091	0.091	0.015	0.091	Significant	Not Significant	Not Significant	Significant	Not Significant
MLP	0.000	0.001	0.001	0.000	0.000	Significant	Significant	Significant	Significant	Significant
XGBoost	0.001	0.015	0.015	0.001	0.001	Significant	Significant	Significant	Significant	Significant
Hybrid LSTM+ TFT	0.999	0.999	0.999	0.965	0.965	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
Voting Regressor	0.336	0.974	0.965	0.336	0.336	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
Voting (LSTM Feature)	0.965	0.999	0.999	0.999	0.999	Not Significant	Not Significant	Not Significant	Not Significant	Not Significant
Voting (TFT Feature)	0.091	0.726	0.726	0.091	0.015	Not Significant	Not Significant	Not Significant	Not Significant	Significant

Note: Bolded results or values marked as "Significant" indicate that the difference between the TFT model and the respective model is statistically significant based on a p-value threshold of $p < 0.05$. Values with $p \geq 0.05$ are considered "Not Significant".

The results indicate that TFT significantly outperforms the MLP and XGBoost models in all five metrics (MSE, RMSE, MAE, MAPE, and R^2), with p-values below the threshold of 0.05. Furthermore, TFT also exhibits statistically significant superiority over the Random Forest model in terms of MSE and MAPE. In contrast, no statistically significant differences were observed between TFT and the LSTM, Hybrid LSTM+TFT, Voting Regressor, or Voting (LSTM Feature) models, indicating comparable performance levels. Notably, although the comparison with the Voting (TFT Feature) model yielded a significant difference only in R^2, all other metrics remained statistically non-significant. These findings suggest that although TFT provides a consistent and competitive performance advantage over traditional and tree-based models, its performance remains on par with that of advanced ensemble and hybrid architectures, particularly those that leverage DL-based feature representations.

3.2 Future Forecasting with TFT Model

The future values of the independent variables employed in the model training are also required to forecast the future 12-month values of the dependent variable (budget expenditures) using the TFT model. These independent variables play an important role in predicting the dependent variable, and the correct prediction of future values increases the prediction accuracy of the dependent variable. The TFT model has a flexible structure that can handle multiple time series data and various types of variables. Hence, it is used to predict the future values of the independent and dependent variables. The current TFT model is trained to forecast both the dependent variable and the independent variables. During the training process, the model learns the relationships of past independent and dependent variables. The future values of the independent variables were predicted using the trained TFT model. The forecasting results of the TFT model for the future 12-month values of the independent variables are shown in Table 7.

Table 7. Future 12-month values of independent variables estimated with the TFT model

Date	Personnel Costs	SSI State Premium Expenses	Goods and Service Purchases	Current Transfers	Capital Expenditures	Capital Transfers	Lending	Interest Expenses
2024-6	237.749	26.867	57.515	333.163	69.121	64.717	20.507	112.584
2024-7	253.253	28.188	59.152	348.503	72.226	68.867	21.987	118.097
2024-8	268.758	29.481	60.789	363.842	75.33	73.017	22.511	123.609
2024-9	284.262	30.767	62.426	379.182	78.435	77.167	22.893	129.122
2024-10	299.766	32.052	64.063	394.522	81.54	81.317	23.252	134.635
2024-11	315.27	33.336	65.751	409.861	84.644	85.468	23.609	140.147
2024-12	330.774	34.621	67.337	425.201	87.749	89.618	23.965	145.66
2025-1	346.279	35.905	68.974	440.541	90.854	93.768	24.321	151.173
2025-2	361.783	37.189	70.611	455.882	93.958	97.918	24.678	156.685
2025-3	377.287	38.473	72.248	471.221	97.063	102.068	25.034	162.198
2025-4	392.791	39.757	73.885	486.559	100.168	106.219	25.39	167.711
2025-5	408.296	41.041	75.522	501.899	103.272	110.369	25.746	173.223

Using these estimated values of the independent variables shown in Table 7, the future values of the dependent variable were estimated. The prediction results for the obtained budget expenditure values are shown in Table 8.

Table 8. Budget expenditures estimated with the TFT model

Date	Future Values (Million TL)
2024-6	859.432
2024-7	888.98
2024-8	891.057
2024-9	913.87
2024-10	988.596
2024-11	1033.263
2024-12	1280.572
2025-1	1327.885
2025-2	1375.199
2025-3	1422.513
2025-4	1469.826
2025-5	1517.14

Table 8 displays the monthly budget expenditure forecasts for the period between June 2024 and May 2025, as predicted by the TFT model, which was identified as the best-performing model in our comparative evaluation. The forecasts demonstrate a gradually increasing expenditure trend, with a notable acceleration beginning in the last quarter of 2024. This projection aligns with seasonal fiscal dynamics and known structural budget cycles observed in the historical data.

4. Conclusion and policy recommendations

This study makes both theoretical and practical contributions by applying a comprehensive set of ML, DL, hybrid and ensemble techniques to forecast Türkiye's public budget expenditures. In addition to individual models such as TFT, XGBoost, MLP, LSTM, and Random Forest, this study introduces new hybrid frameworks and ensemble methods to improve forecast robustness. These model combinations are specifically designed to capture the complex temporal, non-linear, and structural patterns inherent in financial data. The comparative results reveal that although TFT performs exceptionally well, the hybrid and ensemble configurations also exhibit strong predictive capabilities and offer a more holistic understanding of model applicability under various data scenarios. This multifaceted approach provides deeper insights for data-driven public finance management and highlights the transformative potential of AI-based forecasting in diverse enterprise environments.

The data used in this study for the period from January 2008 to May 2024 covers several major global and local events, which has significantly increasing the pattern complexity in the data. In particular, the 2008 global financial crisis was a factor that complicated economic forecasts at the beginning of this period. Then, the European debt crisis in 2009, followed by Türkiye's domestic political volatility and political events such as the Gezi Park events and the attempted coup in 2016, caused economic instability. Moreover, the COVID-19 pandemic, which began in 2020, had a profound impact on the global economy and brought about unexpected changes in government spending. Finally, the Kahramanmaraş earthquake in February 2023 had several significant effects on the Turkish economy. The Presidential Strategy and Budget Directorate estimated the total cost of the earthquake to the Turkish economy to be approximately $103.6 billion (2 trillion Turkish Liras), which was equivalent to approximately 9% of Türkiye's national income in 2023. The increased need for goods and services following the devastation caused by the earthquake has put upward pressure on inflation. In addition, more than 500,000 buildings were damaged in the earthquake zones, and communication and energy infrastructure were severely damaged. This increased reconstruction costs and indirect economic impacts. The earthquake-affected regions account for 8.5% of Türkiye's exports and 6.7% of its imports. Although exports suffered losses in the aftermath of the earthquake, imports also increased, leading to an additional current account deficit in foreign trade. These developments have made budget expenditure forecasting difficult. Nevertheless, the TFT model worked successfully on datasets covering these extraordinary periods. The model has the capacity to handle the complex nature of time series data, allowing it to provide highly accurate forecasts despite the effects of sudden and unexpected changes. This shows that the TFT model can perform robustly even in the face of exogenous shocks, such as economic and political fluctuations emphasizing its pontential value for public finance management and budget planning.

When comparing our results with those of similar studies in the literature, although prior studies have generally applied simpler model, this study is particularly notable for the high processing capacity of the TFT model. For instance, Hájek and Olej (2010) obtained better results using neural networks and SVM than linear regression models. In this study, the TFT model had lower error rates and higher R² values (0.993) owing to its advanced time series data processing capability. Although Noor et al. (2022) obtained high accuracy rates in Malaysia using the FFNN model, the TFT model in our study provides metric results that further improve their findings. Furthermore, Valle-Cruz et al. (2022) performed well in a context of limited public resources using the Random Forest model; however, the TFT model in our study outperforms Random Forest and other models to better capture seasonal and trend changes. Nevertheless, although the PCA-SVR model used by Yu (2024) is effective against seasonal and short-term changes, the TFT model in our study achieves much lower error rates with a MAE of 0.005% and a MAPE of 0.0658%, indicating that it can handle sudden economic changes more successfully. The GRU model implemented by Yang et al. (2023) also had remarkable performance metrics, the TFT model in our study exceeds these results with a MAPE of 0.0813%, thus demonstrating its superiority in processing time series data. While the study from Larson and Overton (2024) on sales tax revenue forecasts emphasizes the importance of data preprocessing and model optimization, our study achieves superior results on more complex datasets using similar techniques. These comparisons show that our study obtains more accurate and reliable forecasts by using more complex and diverse models than prior studies, and that it extends the potential of ML and DL techniques further, especially in public finance forecasting. The most important novel aspect of this study is that it is the first detailed forecasting study in the field of public finance in Türkiye that comprehensively evaluates advanced AI techniques, particularly using the TFT model. This study demonstrates how this model can process complex time series data and make highly accurate forecasts in the process. This is a significant departure from previous research in the literature, which has generally tested less complex models. Moreover, a comparative performance analysis of multiple ML and DL models is instructive for industrial applications.

The findings of this study reveal the potential use of ML and DL techniques for Türkiye's budget expenditure forecasts. These results provide important insights for fiscal planning and policymaking processes. Evaluating the practical applicability of these AI-based forecasting tools requires consideration of the diverse fiscal governance structures across countries. The models developed in this study, particularly those capable of handling high-dimensional, volatile time series, are inherently flexible and can be adapted to both centralized and decentralized budget systems. In centralized settings, such as Türkiye, direct integration into national budget planning offices can improve top-down fiscal control and forecasting accuracy. Conversely, in federal or decentralized systems, the same models can be customized for regional or municipal authorities to support localized expenditure forecasting, thereby enhancing resource allocation autonomy while maintaining macro-level consistency. This cross-system adaptability ensures that the proposed models are not bound to a single institutional framework, but can instead complement various budgetary architectures depending on governance needs and data availability. In addition to their architectural adaptability, the institutional roles of these models must also be clarified. Notably, the aim of this study is not to replace fiscal decision-makers with automated systems, but to demonstrate how AI-based models can be effectively integrated into the decision-making process as advisory tools. These models, particularly interpretable architectures such as TFT, offer explainable and data-driven forecasts that can complement rather than override human judgment. In politically sensitive environments, where fiscal strategies must account for societal priorities, legal frameworks, and dynamic macroeconomic pressures, the role of AI should remain supportive. By highlighting patterns, detecting anomalies, and improving the timeliness of budget insights, these tools can enhance the analytical capacity of policymakers while preserving the necessary discretion and contextual evaluation inherent in public finance management.

Several specific policy implications are indentified based on this adaptive potential. First, more accurate budget forecasts allow the government to maintain fiscal discipline and allocate resources efficiently, while providing a strategic advantage in managing unexpected fiscal burdens and crises periods. With these forecasts, fiscal planners can anticipate future budget deficits and make necessary adjustments. They can also provide guidance for reconsidering government spending, reducing unnecessary investments, and prioritizing investments, especially during periods of projected high expenditures. To maintain economic stability, it is recommended that tax policies and public expenditures be adjusted in line with forecasted budget expenditures. These technologies can also contribute to corruption prevention by increasing transparency and accountability in public financial management. Finally, the continuous integration of these techniques can facilitate a more scientific and data-driven approach to future fiscal planning and budgeting processes, which will support a more efficient use of public resources and help achieve economic development objectives.

Although this study provides important findings, its limitations should be considered. Extending the time span and coverage of the datasets used would allow us to better test the generalizability of the models. Moreover, increasing the number of independent variables considered during model training could further improve the forecasting success. Future work should broaden the base of this research and further explore the potential of ML and DL models in budget forecasting. It would be useful to test the generalizability and robustness of the models using larger and more diversified datasets. This could be critical for assessing model performances under different economic conditions and during periods of political crisis. Moreover, further tuning of the model hyperparameters and comparative testing of different model architectures could improve forecasting accuracy. In addition to the integration of AI-based models into budget forecasting processes, the adaptation of these technologies to other areas of financial management should be examined. For instance, the impact of these models on other financial indicators such as public expenditure analysis and revenue forecasts could be investigated.

In parallel with the technical and institutional advances proposed above, ensuring sustainable implementation will require targeted investments in human capital. Considering the increasing relevance of AI-based forecasting in public finance, capacity-building efforts must align with institutional needs. Accordingly, we propose the implementation of structured training programs and workshops specifically tailored for policymakers, budget officers, and public-sector data analysts. These programs should be designed to cover not only the technical aspects of AI models (e.g., model interpretability, data preprocessing, and forecasting logic), but also practical integration strategies within existing budget planning workflows. Content modules could include hands-on model usage, ethics of algorithmic decision-making, and real-time policy simulations. Training can be delivered through a hybrid model that combines in-person sessions (e.g., within national audit institutions or budget directorates) with asynchronous online courses supported by academic or multilateral institutions to ensure accessibility and scalability. Collaboration with universities, public finance schools, and organizations such as the IMF, OECD, or World Bank could help standardize training curricula and provide international benchmarking. Case studies from countries such as India and Brazil, where similar initiatives have been successfully piloted, can provide as reference frameworks. By grounding the training in both technical skill-building and policy relevance, such programs can accelerate the institutional readiness for AI adoption in public financial management. These recommendations can help expand the role of AI in public financial management and maximize the benefits of these technologies.

References

Abtew, A., Demissie, D., & Kekeba, K. (2023). An Ontology-Driven Machine Learning Applications for Public Policy Analysis from Social Media Data: A Systematic Literature Review. J Curr Trends Comp Sci Res, 2(2), 182-190. https://doi.org/10.33140/jctcsr.02.02.13

Ajiga, D. I., Adeleye, R. A., Asuzu, O. F., Owolabi, O. R., Bello, B. G., & Ndubuisi, N. L. (2024). Review of ai techniques in financial forecasting: applications in stock market analysis. Finance & Accounting Research Journal, 6(2), 125-145. https://doi.org/10.51594/farj.v6i2.784

Akkaya, M. (2022). Triple deficit pressure index and estimation of the financial crisis: the case of Turkey. Mehmet Akif Ersoy University Journal of Faculty of Economics and Administrative Sciences, 9(3), 1507-1521. https://doi.org/10.30798/makuiibf.802005

Ali, M., Imran, M., Hamza, M., Jehanzeb, M., Habib, S., & Sajid, M. (2019). Industrial financial forecasting using long short-term memory recurrent neural networks. International Journal of Advanced Computer Science and Applications, 10(4). https://doi.org/10.14569/ijacsa.2019.0100410

Aliu, B. (2019). Big data phenomenon in banking. Texila International Journal of Academic Research, 6(2), 81-87. https://doi.org/10.21522/tijar.2014.06.02.art008

Alohali, M. A., El-Rashidy, N., Alaklabi, S., Elmannai, H., Alharbi, S., & Saleh, H. (2024). Swin-ga-rf: genetic algorithm-based swin transformer and random forest for enhancing cervical cancer classification. Frontiers in Oncology, 14:1392301. https://doi.org/10.3389/fonc.2024.1392301

Asriani, A., Rianse, U., Surni, S., Taufik, Y., & Herdhiansyah, D. (2023). Forecasting model of corn commodity productivity in indonesia: production and operations management, quantitative method (pom-qm) software. International Journal of Advanced Computer Science and Applications, 14(5). https://doi.org/10.14569/ijacsa.2023.0140565

Bağdigen, M. (2002). How Accurate is Revenue Forecasting in Turkey? An Empirical Analysis. Yapı Kredi Economic Review, 29-37.

Bhattacharjee, R., Gupta, A., Das, N., Agnihotri, A. K., Ohri, A., & Gaur, S. (2022). Analysis of algal bloom intensification in mid-ganga river, india, using satellite data and neural network techniques. Environmental Monitoring and Assessment, 194(8). https://doi.org/10.1007/s10661-022-10213-6

Cai, X., Yuan, W., Liu, X., Wang, X., Chen, Y., Deng, X., Wu, Q., Han, K., Cao, Z., Wu, W., & Wang, B. (2022). Deep Learning Model for Soil Environment Quality Classification of Pu-erh Tea. Forests, 13(11), 1778. https://doi.org/10.3390/f13111778

Chen, T. and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. https://doi.org/10.48550/arxiv.1603.02754

Cui, Q., Rong, S., & Zhang, B. (2023). Advancing the comprehension of consumer price index and influencing factors: insight into the mechanism based on prediction machine learning models. Advances in Economics and Management Research, 7(1), 125-128. https://doi.org/10.56028/aemr.7.1.125.2023

Çınaroğlu, S. and Başer, O. (2019). Does the unification of health financing affect the distribution pattern of out-of-pocket health expenses in Turkey?. International Journal of Social Welfare, 28(3), 293-306. https://doi.org/10.1111/ijsw.12389

Devarajan, S., Swaroop, V., & Zou, H. (1996). The composition of public expenditure and economic growth. Journal of Monetary Economics, 37(2-3), 313-344. https://doi.org/10.1016/0304-3932(96)01249-4

Dzigbede, K. D., Pathak, R., & Muzata, S. (2022). Budget systems and post-pandemic economic resilience in developing countries. Journal of Public Budgeting, Accounting & Financial Management, 35(3), 333-353. https://doi.org/10.1108/jpbafm-03-2021-0036

Eryılmaz, F. and Murat, D. (2016). Searching for political business cycles in turkey: findings from fiscal policy. International Journal of Economic and Administrative Studies, 0(17), 197. https://doi.org/10.18092/ijeas.11309

Gachhadar, P. K., Baniya, C. B., & Mandal, T. N. (2024). Pattern of plant biomass and carbon stock along different elevational forests in eastern Nepal. *Banko Janakari*, *34*(1), 15–29. https://doi.org/10.3126/banko.v34i1.62716

HairJr., J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). Evaluation of the structural model. Classroom Companion: Business, 115-138. https://doi.org/10.1007/978-3-030-80519-7_6

Hájek, P. and Olej, V. (2010). Municipal revenue prediction by ensembles of neural networks and support vector machines. WSEAS Transactions on Computers, 9(11), 1255-1264.

Hertel, L. (2020). Quantity vs. quality: on hyperparameter optimization for deep reinforcement learning. https://doi.org/10.48550/arxiv.2007.14604

Hossain, M., Ismail, M., & Hossain, M. (2022). Enhancing stock price prediction using empirical mode decomposition, rolling forecast and combining statistical methods. International Journal of Computing and Digital Systems, 12(6), 1343-1356. https://doi.org/10.12785/ijcds/1201108

Kairu, A., Orangi, S., Mbuthia, B., Ondera, J., Ravishankar, N., & Barasa, E. (2021). Examining health facility financing in kenya in the context of devolution. BMC Health Services Research, 21(1). https://doi.org/10.1186/s12913-021-07123-7

Kanupriya. (2024). Opportunities and challenges of AI in financial risk management: A brief labour-centric analysis. *Forum for Economic and Financial Studies*, *2*(2), 1758. https://doi.org/10.59400/fefs1758

Kara, B. (2024a). The Performance of Medium-Term Budgeting in Türkiye: An Analysis of Budget Forecasts. *Journal of Management and Economics*, *31*(4), 659-676. https://doi.org/10.18657/yonveek.1521261

Kara, B. (2024b). The Impact of Budget Revenue and Expenditure Forecasting Errors on Inflation in Turkey: An Analysis of the 1975-2021 Period. *İstatistik ve Uygulamalı Bilimler Dergisi*, (9), 19-28. https://doi.org/10.52693/jsas.1417919

Köse, N., and Ünal, E. (2024). The Roles of the Terms of Trade and the Real Exchange Rate in the Current Account Balance. *Economics*, *18*(1), 20220117. https://doi.org/10.1515/econ-2022-0117

Kuzheliev, M., Рекуненко, І. І., Нечипоренко, А. В., & Nemsadze, G. (2019). Discretionary budget expenditure in the system of state regulation of the country's socioeconomic development. Public and Municipal Finance, 7(4), 8-18. https://doi.org/10.21511/pmf.07(4).2018.02

Laborda, J., Ruano, S., & Zamanillo, I. (2023). Multi-country and multi-horizon gdp forecasting using temporal fusion transformers.. https://doi.org/10.20944/preprints202305.0445.v1

Larson, S. and Overton, M. (2024). Modeling approach matters, but not as much as preprocessing: Comparison of machine learning and traditional revenue forecasting techniques. Public Finance Journal, 1(1), 29-48. https://doi.org/10.59469/pfj.2024.8

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2019). Temporal fusion transformers for interpretable multi-horizon time series forecasting.. https://doi.org/10.48550/arxiv.1912.09363

Lin, S. and Huang, H. (2020). Improving deep learning for forecasting accuracy in financial data. Discrete Dynamics in Nature and Society, 2020, 1-12. https://doi.org/10.1155/2020/5803407

Liu, J. and Liu, J. (2022). Permeability predictions for tight sandstone reservoir using explainable machine learning and particle swarm optimization. Geofluids, 2022, 1-15. https://doi.org/10.1155/2022/2263329

Liu, P. (2024). Time series analysis and algorithm for fluctuation pattern recognition and forecasting of economic indicators. jes, 20(6s), 812-816. https://doi.org/10.52783/jes.2746

Maeda, I., Matsushima, H., Sakaji, H., Izumi, K., deGraw, D., Kato, A., & Kitano, M. (2021). Predictive uncertainty in neural network-based financial market forecasting. International Journal of Smart Computing and Artificial Intelligence, 5(1), 1-18. https://doi.org/10.52731/ijscai.v5.i1.541

Naumoski, A., Upadhya, M., & Zdraveva, P. (2022). Estimating public climate finance using objective– based cost component approach. Economic Development, 24(4), 126-147. https://doi.org/10.55302/ed22244126n

Ngoc, T. T., Đại, L. V., & Minh, L. B. (2022). Effects of data standardization on hyperparameter optimization with the grid search algorithm based on deep learning: a case study of electric load forecasting. Advances in Technology Innovation, 7(4), 258-269. https://doi.org/10.46604/aiti.2022.9227

Noor, N., Sarlan, A., & Aziz, N. (2022, February). Revenue Prediction for Malaysian Federal Government Using Machine Learning Technique. In Proceedings of the 2022 11th International Conference on Software and Computer Applications (pp. 143-148). https://doi.org/10.1145/3524304.3524337

Obadić, A., Globan, T. & Nadoveza, O. (2014) Contradicting the Twin Deficits Hypothesis: The Role of Tax Revenues Composition, Panoeconomicus, (6): 653-667. https://doi.org/10.2298/PAN1406653O

Olubusola, O., Mhlongo, N. Z., Daraojimba, D. O., Ajayi-Nifise, A. O., & Falaiye, T. (2024). Machine learning in financial forecasting: a u.s. review: exploring the advancements, challenges, and implications of ai-driven predictions in financial markets. World Journal of Advanced Research and Reviews, 21(2), 1969-1984. https://doi.org/10.30574/wjarr.2024.21.2.0444

Önal, D. K. (2024) The Buchanan-Wagner Hypothesis: Revisiting the Theory with New Empirics for a Spendthrift Democracy, Panoeconomicus, 71(3): 433-454. https://doi.org/10.2298/PAN200522009K

Özcan, S. (2017). Evaluation of the Accuracy of General Budget Revenue and Expenditure Estimates between 1924-2012 in Turkey. Gazi University Journal of Faculty of Economics and Administrative Sciences, 19(2), 701-724.

Özcan, S. and Günal, T. (2024) Testing the Triple Deficits in the Emerging Economies of Europe, Panoeconomicus, 1–17. https://doi.org/10.2298/PAN201206017O

Özcan, S., and Tosun, M. (2014). Evaluation of the Ministry of National Education budget forecasts in terms of accuracy principle. Socioeconomics, 22(22).

Özekicioğlu, S. S. and Tülümce, S. Y. (2020). The impacts of corruption on budget balance and public debt in turkey: an empirical analysis. Journal of Management and Economics Research, 18(3), 46-60. https://doi.org/10.11611/yead.775529

Özker, A. N. (2024). Financial performance objective of public expenditures in last period and it's in connected with economic budget in Turkey. *South Florida Journal of Development*, *5*(9), e4431. https://doi.org/10.46932/sfjdv5n9-055

Parlak, M. (2005). Exceeding Expenditure Estimates in the Budget and Its Economic Effects. Journal of Court of Accounts, 73-87.

Patcharaprakiti, N., Kirtikara, K., Jivacate, C., Sangswang, A., Tunlasakun, K., & Muenpinij, B. (2010). System identification with cross validation technique for modeling inverter of photovoltaic system.. https://doi.org/10.1109/icmet.2010.5598430

Quteishat, E. M. A. (2024). Predictive modelling in legal decision-making: leveraging machine learning for forecasting legal outcomes. Journal of Electrical Systems, 20(3), 2060-2071. https://doi.org/10.52783/jes.4006

Republic of Türkiye Ministry of Treasury and Finance, (2024), Monthly Budget Realization Reports, https://www.hmb.gov.tr/bumko-aylik-butce-gerceklesme-raporlari, Access Date: 04.06.2024

Robinson, M. (1998). Measuring compliance with the golden rule. Fiscal Studies, 19(4), 447-462. https://doi.org/10.1111/j.1475-5890.1998.tb00295.x

Sahu, A. and Pattnaik, S. (2017). Feature selection using evolutionary functional link neural network for classification. International Journal of Advances in Applied Sciences, 6(4), 359. https://doi.org/10.11591/ijaas.v6.i4.pp359-367

Shantal, M. (2023). A novel approach for data feature weighting using correlation coefficients and min–max normalization. Symmetry, 15(12), 2185. https://doi.org/10.3390/sym15122185

Shao, Z., Zhang, Z., Wang, F., & Xu, Y. (2022). Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting.. https://doi.org/10.48550/arxiv.2206.09113

Shkolnyk, I., Ladyka, Y., Orlov, V., Aldiwani, K., & Kozmenko, Y. (2021). Balancing state budget expenditures: a case of ukraine using the simplex method. Public and Municipal Finance, 10(1), 34-46. https://doi.org/10.21511/pmf.10(1).2021.04

Sinshaw, N. T., Ejigu, B. E., Assefa, B. G., & Mohapatra, S. K. (2023). Amharic handwritten & machine printed character recognition using deep cnn with random search hyperparameter optimization algorithm. https://doi.org/10.21203/rs.3.rs-2878655/v1

Šuliková, V., Siničáková, M. & Horváth, D. (2014) Twin Deficits in Small Open Baltic Economies, Panoeconomicus, 2: 227-239. https://doi.org/10.2298/PAN1402227S

Sun, D. (2015). Book review: state government budget stabilization: policy, tools, and impacts. The American Review of Public Administration, 45(2), 241-242. https://doi.org/10.1177/0275074014545487

Şenesen, G. (2000). Quantitative approaches to the evaluation of the performance of initial budget appropriations. 15th Turkish Finance Symposium, (pp. 345-375). Antalya.

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Rojas-Hernández, R. (2022, September). Public Budget Simulations with Machine Learning and Synthetic Data: Some Challenges and Lessons from the Mexican Case. In International Conference on Electronic

Governance with Emerging Technologies (pp. 141-160). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22950-3_12.

Wu, B., Wang, L., Tao, R., & Zeng, Y. (2022). Interpretable tourism volume forecasting with multivariate time series under the impact of covid-19. Neural Computing and Applications, 35(7), 5437-5463. https://doi.org/10.1007/s00521-022-07967-y

Yang, C. H., Molefyane, T., & Lin, Y. D. (2023). The Forecasting of a Leading Country's Government Expenditure Using a Recurrent Neural Network with a Gated Recurrent Unit. Mathematics, 11(14), 3085.

Yardım, M. S., Çilingiroğlu, N., & Yardım, N. (2013). Financial protection in health in turkey: the effects of the health transformation programme. Health Policy and Planning, 29(2), 177-192. https://doi.org/10.1093/heapol/czt002

Yaşa, A. A., Şanlısoy, S., & Özen, A. (2020). Bütçe Tutarlılığı ile Politik İstikrarsızlık İlişkisi: Türkiye'de 1984-2018 Dönemi Analizi. *Sosyoekonomi*, *28*(44), 337-354.

Yılmaz, H. H. (2003). Konsolide Bütçe Gelir ve Gider Tahminlerinin Gerçekleşmelere Göre Güvenilirlik Düzeyi. International Conference in Economics IV METU-ERC. Ankara.

Yu, N. and Haskins, T. (2021). Knn, an underestimated model for regional rainfall forecasting. https://doi.org/10.48550/arxiv.2103.15235

Yu, X. (2024). Construction of Local Fiscal Revenue Forecasting Model Based on SVR Model. Applied Mathematics and Nonlinear Sciences, 9(1), 1-12. https://doi.org/10.2478/amns-2024-0784.

Yun, D., Yang, H., Kim, S. G., Kim, K., Kim, D. K., Oh, K-H., Joo, K. W., Kim, Y. S. & Han, S. S. (2023). Real-time dual prediction of intradialytic hypotension and hypertension using an explainable deep learning model. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-45282-1

Zakaria, S., Manaf, S. M. A., Amron, M. T., & Suffian, M. T. M. (2023). Has the world of finance changed? a review of the influence of artificial intelligence on financial management studies. Information Management and Business Review, 15(4(SI)I), 420-432. https://doi.org/10.22610/imbr.v15i4(si)i.3617

Zatonatska, T., Liashenko, O., Fareniuk, Y., Dluhopolskyi, O., Dmowski, A., & Cichorzewska, M. (2022). The migration influence on the forecasting of health care budget expenditures in the direction of sustainability: case of ukraine. Sustainability, 14(21), 14501. https://doi.org/10.3390/su142114501

Zhou, H., Xu, K., Bao, Q., Lou, Y., & Qian, W. (2024). Application of conversational intelligent reporting system based on artificial intelligence and large language models. Journal of Theory and Practice of Engineering Science, 4(03), 176-182. https://doi.org/10.53469/jtpes.2024.04(03).16