Critical Study of Machine Learning Models for Retail Sales Prediction

Research Practicum Part 2
MSc Data Analytics

Athul Ajayghosh .
Student ID: 23416068

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Athul Ajayghosh .
Student ID:	23416068
Programme:	MSc Data Analytics **Year:** 2025
Module:	Research Practicum Part 2
Supervisor:	Prof.Vikas Sahni
Submission Due Date:	Thursday 11th December 2025
Project Title:	Critical study of Machine Learning Models for Retail Sales Prediction
Word Count:	8571 **Page Count:** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……………………………………………………………………………………………………

Date: 10-12-2025 ……………………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Critical Study of Machine Learning Models for Retail Sales Prediction

Athul Ajayghosh .
23416068

Abstract

Proper sales projections of retailing is essential in minimizing inventory expenses and enhancing efficiency. The traditional models like the ARIMA and SARIMA have been found to be useful in predicting the short run but they do not cope with the non-linearity and dynamism of the current retail data. The latest developments in the field of forecasting have presented machine learning algorithms such as Random Forest, XGBoost and CatBoost, as well as deep learning based algorithms such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and hybrid Models. More recent scalable models shows good long-horizon accuracy, but there are few studies which report their performance in terms of accuracy, efficiency, and reduction of inventory costs in retail sales forecasting.

The paper compares fourteen forecasting models of the weekly revenue of the UCI Online Retail data based on a constant four-week prediction horizon. The classical models had moderate accuracy (ARIMA RMSE 63796; SARIMA 57585), whereas the machine learning models enhanced their accuracy by non-linear learning of features. The best results were obtained with the method of deep learning, and the lowest error was obtained with CNN-LSTM (RMSE 31948; MAPE 8.2%).These results underscore the usefulness of sophisticated neural models in enhancing sales prediction and minimizing inventory-related expenses and offer retailers some practical ideas and empirical data to the current forecasting literature.

1 Introduction

1.1 Background

Retail operations take place through correct sales forecasting. It supports the inventory management, logistics, pricing, and promotional planning which have a direct impact on profitability and customer satisfaction. Inaccuracies in forecasting may put a lot of financial strain, overestimation may result in overstocking, increased holding costs, and lost capital, whereas underestimation may cause stocks out and missed sales opportunities. In turn, high accuracy of the forecasts has become the priority of retailers that aim to maximise the operational efficiency and minimise the costs of the inventory (Pathak et al., 2025; Dong and Hao, 2024). It is consistent with other studies showing that a clustering and segmentation method can identify the trends of the operation in the retail setting (Arbsuwan, 2023) and that retail sales data can be directly used to aid the warehouse planning and optimalisation of operations (Arbsuwan and Buddhakulsomsiri, 2023).

There are various phases in the development of forecasting models in history. The initial methods, including Autoregressive Integrated Moving Average (ARIMA) and seasonal variant of ARIMA, were used in time-series analysis over decades, as they are easy to interpret, they are statistically significant, and they perform well in the short-term (Hyndman and Khandakar, 2008; Azzahra and Sudirman, 2025).With the growth of computational capacity and availability of data, machine learning algorithms like Random Forest , XGBoost and CatBoost have become strong alternatives. These frameworks are able to model complex, non-linear relationships and interactions among a number of input variables, and are thus more flexible in the dynamic retail environment modelling (Raj et al., 2025; Sankar et al., 2025). In a similar way, deep learning (DL) models especially Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are especially effective to model long range temporal dependencies and sequencing relations in time-series data (Garcia et al., 2025; Mansur et al., 2025). CNN-LSTM is also a hybrid architecture that better feature extractor, which integrates convolutional layers (short-term local patterns) with recurrent layers (long-term dependencies) (Zhou et al., 2024).

This trend in the development of forecasting algorithms also raises a research gap since most models have shown excellent results in the context of individual models, but very few studies have compared classical, machine learning, deep learning, and hybrid models in the same experimental framework particularly in the context of retail applications geared towards cutting the inventory cost.

In order to fill in this gap, the paper applies the UCI Online Retail dataset to conduct a comprehensive comparative analysis of 14 forecasting models, including statistical, machine learning, deep learning, and hybrid paradigms. The data set gives a realistic and high frequency picture of the transactional retail behavior and has the capability to provide strong performance analysis of the models both in terms of accuracy and computation efficiency.

1.2 Importance and Motivation

Inventory is still one of the biggest cost drivers of retail business. Overstocking is a waste of capital and higher storage costs, whereas understocking is lost sales and dissatisfied customers (Dong and Hao, 2024; Jing et al., 2024). Effective and precise forecasting directly reduces these risks through having the level of inventory matching the anticipated demand.

Conventional approaches such as ARIMA and SARIMA are easy to compute and understand but tend to fail to detect non linear behaviour and external variables. On the other hand, more advanced models like XGBoost, GRU, and hybrid CNN-LSTM structure are able to analyze the intricate temporal dynamics and get more accurate predictions, but has a higher demand on computing resources (Sankar et al., 2025; Mansur et al., 2025). Attention-based and distributed algorithms such as AttnBoost and DARIMA are solutions to these trade-offs, in that they can be more interpretable, scaled larger, and with better accuracy.

The concept behind this study is to find the means of forecasting that which gives the desired accuracy, efficiency and applicability in retail stores that directly aids in the decision-making and reduction of inventory and operational costs.

1.3 Research Question and Objectives

The research question which is studied:

"How effective are classical, machine learning, deep learning, and hybrid forecasting model such as ARIMA, SARIMA, Random Forest, XGBoost, LSTM, GRU, CNN–LSTM, and AttnBoost in forecasting retail sales trends and reducing inventory related costs?"

1. **Implement** a large variety of forecasting models, such as statistical , machine learning , deep learning, and hybrid models.
2. **Evaluate** forecasting performance on the based accuracy measures of RMSE, MAE, and MAPE.
3. **Measure** computational efficiency based on training time, prediction time, memory usage and throughput.
4. **Analyse** the predictive ability of each model in the reduction of inventory costs and operational decision-making.

Through these goals this study determines the best forecasting methodology which provides a balance between predictive accuracy, computing power.
This study helps to research on the performances of the traditional and modern forecasting methods under equal circumstances of retail. It offers some insights that can assist retail companies and data analysts to select the most suitable forecasting technique in enhancing sales planning and decision making.

1.4 Limitations and Assumptions

The current study is done based on the UCI Online Retail data[1], which, though detailed, is data on one organisation and might not be generalised to multi-store or multi-regional retailing. The study is based on the assumption that the consumer buying behaviour does not change in the forecast horizon and any external disrupting factor like the economic shocks is insignificant. Complexity of explored deep learning architectures might also be restricted by computational constraints. However, these parameterised parameters guarantee comparisons of models fairly and reproducibly.

1.5 Structure of the Report

The rest of this dissertation is structured in the following way:

- **Section 2** - Related Work: Reviews the literature that has been conducted on classical, machine learning, deep learning and hybrid forecasting models used in retail sales. It provides a critical analysis of past research, their strengths and weaknesses as well as the gaps that this research seeks to fill.
- **Section 3** - Research Methodology: The general research scheme, data gathering, data preprocessing strategies, and assessment measurements. It also describes the experimental design of model training, testing and comparison.
- **Section 4** - Design Specification: Shows the system design, data flow and model design of each forecasting method. This part describes feature engineering, parameter choice and model configuration plans.
- **Section 5** - Implementation: Practically describes the process of coding and implementation in Python in Jupyter Notebook, and includes code libraries, data piping and training configurations of all models.
- **Section 6** - Evaluation: Reports the results of the experiment, comparing the performance of the all forecasting models in terms of accuracy, computational performance and scalability. It has diagrams and debates about the effects of the accuracy of forecast of reducing inventory costs.

[1] https://archive.ics.uci.edu/dataset/352/online%2Bretail

- **Section 7** - Conclusion and Future Work: Concludes the major findings and gives recommendations regarding future research.

2 Related Work

Retail sales forecasting is crucial in the optimisation of inventory management, pricing, and operational planning. Proper forecasting allows the retailers to keep their stocks at the right stock amounts, save money on stocking and also be in a position to react to the changing consumer demand effectively. On the other hand, inefficient forecasting can result in overstocking, stockouts and financial inefficiencies. In the last 20 years, forecasting research has developed, replacing the application of statistical models that are linear and stationary, to machine learning and deep learning models which is able to model non-linear and multivariate dependencies. In the more recent times, there has been the development of hybrid and attention-based architectures to trade off interpretability, accuracy and computational efficiency.

This review critically assesses the available forecasting models in four groupings as classical statistical, machine learning, deep learning, and hybrid/ensemble models and analyses their comparative performance, weaknesses, and contributions. It concludes by outlining some of the major research gaps, which warrant the need to have a detailed comparative study like the one proposed.

2.1 Classical Statistical Approaches

The classical ARIMA and SARIMA (Seasonal ARIMA) models are the most understandable and the most commonly used models in forecasting time-series data. Pathak et al. (2025) have shown that ARIMA works well in stationary manufacturing data, whereas Azzahra and Sudirman (2025) also used SARIMA on daily sales in a coffee shop, which can give good short-term predictive accuracy. Dong and Hao (2024) also integrated ARIMA and linear programming to optimise the pricing and restocking process demonstrating the adaptability of the model. Financial forecasting evidence validates the reliability of ARIMA to predict in a structured and short-term task. It has been shown that in stock markets, ARIMA-based stock price forecasts enhanced option-trading profitability despite volatility issues, and this indicates the power as well as weakness of ARIMA in dynamic settings (Dave and Patel, 2025).

These methods are however linear in nature and assume stationarity, which limits them to non-linear consumer behaviour or other external influences like promotions and holidays. To overcome these problems, Hyndman and Khandakar (2008) proposed Auto-ARIMA, whereby the auto-tuning of parameters is performed based on information criteria, and Wang et al. (2023) proposed Distributed ARIMA (DARIMA) in which the auto-tuning of the parameters is computed at a massive scale (ultra-long series). Although these innovations enhanced automation, and the capability of the computation, they failed to address the fundamental drawback of ARIMA, that it could not model multivariate data or highly volatile retail data.

Critical Evaluation:

The use of statistical models is still necessary in making benchmark comparisons since they are transparent, easy to explain, and can be readily computed. However, they are not flexible enough and do not have adaptive learning processes suitable to the complex non-linear form of modern retail data.

2.2 Machine Learning Approaches to Retail Forecasting

Machine learning brought about a paradigm shift in that, it made the models to learn without having to make rigid statistical assumptions. Random Forest (RF) and gradient boosting algorithms like XGBoost, CatBoost and LightGBM have achieved good performance in various retail applications. Christopher and Rahmatillah (2025) discovered that RF was better in SME digital ecosystems when using exogenous features, including promotions and holidays. Raj et al. (2025) have added sentiment and market-trend analysis to ML models and demonstrated the latter two inputs enhanced forecast performance by a large margin. The current state of regression studies also puts emphasis that most practical time series in the real world should be modelled using a hybrid or non-linear fitting. Clark et al. (2020) say that classical regression models are not adequate to model multidimensional covariate interactions, and they advocate the adoption of dynamic regression, regARIMA and deep learning (Clark et al., 2020) models.

Sankar et al. (2025) performed a large scale comparison on Walmart data over the Random Forest, XGBoost and GRU structure. The experiment did find that GRU and ensemble boosting model had better accuracy, and boosting frameworks provided good trade-offs between accuracy and computation speed. Equally, Jiang, Ruan, and Sun (2024) compared Prophet and LightGBM, concluding that LightGBM delivered lower RMSE values, as well as, improved the capture of the effect of holidays. These results prove that ensemble ML methods are effective to model non-linear dependency and heterogeneous features of retail sales.

Critical Evaluation:

Compared to ARIMA and SARIMA, machine learning models are more accurate especially when dealing with feature-rich data set since they are capable of modeling interactions among multiple explanatory variables. Nevertheless, they need thorough feature engineering, they can easily overfit, and are not always interpretable, which makes their adoption in working retail environments difficult. Large ensemble computational cost is also a problem to real time forecasting.

.

2.3 Deep Learning for Retail Forecasting

Deep learning has been a major breakthrough that has helped the time-series forecasting process through automatic identification of time and contextual patterns on the sequential data. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are able to learn long-term connections which classical models fail to acknowledge. Garcia et al. (2025) made a comparative evaluation of ARIMA and LSTM in forecasting retail financial states and found out that in a dynamic setting, LSTM provided much better results.

Mansur et al. (2025) suggested a hybrid CNN-LSTM network, which integrates convolutional layers, which assist in the recognition of short-term patterns, and LSTM layers, which help in learning temporal patterns. This model had a mean absolute percentage error (MAPE) of 4.16 which is better than single neural networks. Zhou et al. (2024) utilized this concept and combined ARIMA with CNN-LSTM to form a residual-corrected hybrid, which was more accurate than the pure ARIMA by over 20%.

Recent research on the comparison of Prophet and LSTM models in sales prediction of coffee shops NetSales revealed that the latter was able to measure seasonal discontinuities more precisely, opposing the smaller values of MAPE, and thus LSTM is more powerful in modeling complex time variability (Sudirman, Rahmatillah and Herwanto, 2025).New models based on attention mechanism (such as AttnBoost and Temporal Fusion Transformers (TFTs)) have been developed recently, and this dynamically assigns significance to time steps, a feature that not only lead to better predictions but also makes them easier to understand. AttnBoost is a boosting based algorithm that integrates attention and the TFTs are time and covariate based algorithms that use explainable multi-horizon predictions.

Critical Evaluation:

The deep learning methods are good at processing complex and non-linear retail data. They are however, computationally expensive, data consumptive and are generally viewed as black boxes. Despite the enhanced interpretability of attention mechanisms, their direct application and optimization needs a lot of computational skills and resources.

2.4 Hybrid and Ensemble Forecasting Models

It is the hybrid solutions that are attempting to combine the strengths of multiple paradigms. Hybrids of ARIMA-XGBoost and ARIMA-CatBoost (Abdullah et al., 2025; Jiang et al., 2024) employ both linear and non-linear model fits in the sense that trend and seasonality are first modelled using ARIMA and then Boosting models are used to handle the complexities within the residuals. Collectively, these studies demonstrate that there is a clear trend where machine learning and hybrid models are becoming the preferred methods of forecasting retail demand because of their capability to modelling intricate and non-linear behaviours (Christopher and Rahmatillah, 2025; Caglayan-Akay et al., 2025). As was demonstrated by Zhou et al. (2024), the ARIMA-CNN-LSTM model was more accurate due to the combination of statistical and deep learning. Similar benefits were established by Mansur et al. (2025) when CNN-LSTM hybrids were used on multivariate retail sales, the contextual variables of weather and holidays were important determinants of demand. Cross-functional solutions are becoming more popular. As shown by Singh and Gupta (2025), the accuracy of the anomaly detection was significantly higher when reinforcement learning is used together with Random Forests, and it can be argued that this aspect of hybrid architectures shows how much more can be done when complementary models are used to benefit (Singh and Gupta, 2025).

In addition, Prophet (Taylor & Letham, 2018) has been discussed because of its interpretability and simplicity, with automated trend, seasonality, and holiday decomposition effects. Prophet is an effective model used to compare more complicated models because of its transparency and strong ability to deal with non-regular seasonality.

Critical Evaluation:

The hybrid and ensemble model are always more effective than single-model methods in terms of combining the strengths that ARIMA, ML, and DL have interpretability, flexibility, and the ability to learn patterns respectively. Nonetheless, they too acquire complexity of their components, making it more complex to implement and more prone to overfitting. Without standardisation of evaluation frameworks, it is difficult to compare hybrid approaches across studies.

2.5 Research Gaps

Although the forecasting methodologies have evolved at a high rate, there are still critical gaps in the literature. A majority of the research measures models solely based on the accuracy measures (RMSE, MAE, MAPE) but does not consider the efficiency measures like the computational time, home memory, and scalability considerations that are important in real retail implementation. Besides, the studies that have correlated the accuracy of the forecasting with the reduction of inventory cost are few, the most significant working factor of the retailers. Limited comparative literature has compared a diverse range of statistical and deep learning models in a single and consistent experimental design. More so, the akin literature may have been primarily concentrated on domain specific data, thus, generalisation in retail settings is challenging.

The proposed study solves these limitations by deploying and comparing fourteen forecasting models, which include ARIMA, SARIMA, Auto-ARIMA, DARIMA, Random Forest, XGBoost, CatBoost, LightGBM, LSTM, GRU, CNN-LSTM, AttnBoost, ARIMA-XGBoost, ARIMA-CatBoost, ARIMA-CNN-LSTM, and Prophet. Both predictive accuracy and computational efficiency are used to evaluate each model and results are interpreted considering inventory cost reduction and decision support.

Despite these methods delivering impressive predictive scaling, there is limited comparative research on the systematic assessment of the accuracy, computational and implication of operation of the classical, machine learning, deep learning, and hybrid models. Majority of the research is limited to study the accuracy of prediction but does not address runtime efficiency and memory usage and scalability which are the direct factors impacting on the implementation process in retail system. In addition, there are minimal studies that specifically relate the accuracy of forecasting to the reduction of inventory cost which is one of the key goals of the retail operations management.

Such a assessment scenery causes difficulties in making practitioners choose the forecasting methods that are not only accurate but also practical in their application.

Table 1: Comparison of Studies

No.	Study	Dataset / Domain	Models Compared	Evaluation Metrics RMSE,MAE,MAPE,PPS(Predictions/sec)	Key Findings
1	Dong & Hao (2024)	Vegetable retail	ARIMA + Linear Programming	RMSE ✓ MAE ✗ MAPE ✗ PPS ✗	ARIMA(0,1,1)*(0,1,1) achieved strong short-term accuracy; supports pricing optimisation.
2	Raj et al. (2025)	E-commerce retail	ARIMA, Random Forest, LSTM	RMSE ✓ MAE ✗ MAPE ✓ PPS ✗	LSTM and RF surpassed ARIMA in accuracy; inclusion of sentiment data improved performance.
3	García et al. (2025)	Financial retail	ARIMA vs LSTM	RMSE ✓ MAE ✓ MAPE ✗ PPS ✗	LSTM achieved lower RMSE under volatile sales trends; ARIMA remained computationally efficient.
4	Abdullah et al. (2025)	Virtual trade forecasting	ARIMA–CatBoost	RMSE ✓ MAE ✗ MAPE ✗ PPS ✗	Hybrid ARIMA–CatBoost reduced trend bias and improved stability vs standalone ARIMA.
5	Mansur et al. (2025)	Retail stores	CNN–LSTM vs ANN, CNN, LSTM	RMSE ✗ MAE ✗ MAPE ✓ PPS ✗	CNN–LSTM achieved MAPE ≈ 4.16%; contextual variables boosted accuracy.

6	Sankar et al. (2025)	Walmart weekly sales	RF, XGBoost, LSTM, GRU	RMSE ✓ MAE ✓ MAPE ✗ PPS ✗	GRU achieved best accuracy; boosting methods competitive and efficient.
7	Jiang, Ruan & Sun (2024)	Walmart sales	Prophet vs LightGBM	RMSE ✓ MAE ✗ MAPE ✗ PPS ✗	LightGBM produced lower RMSE and superior holiday pattern detection.

3 Research Methodology

The research aims to use a quantitative, experimental approach to research that can be replicated fully. The overall goal of the research is to compare a comparative body of forecasting methods classical statistical models, machine learning models, deep learning architectures, and their combination through the use of purely controlled and transparent conditions. The methodology specifies the specific steps taken, the computing environment, the data preparation flow, and the processes of model construction, and the criteria to evaluate the results. Even though the research loosely relies on the CRISP-DM model as a conceptual framework, the emphasis in this case is on the scientific approach to the execution of a provable forecasting experiment instead of an overall analytics workflow.

The research design of the study was designed in such a way that there was fair and consistent testing of all the forecasting models. To do so, each of the models was trained with the same historical time-series data that was based on weekly aggregated retail revenue and tested using same four weeks forecasting horizon. This same design will ensure there is no chance of the differences in performance due to the variation in data or the dissimilarity in samples. Rather it guarantees that disparity in performance between models can be related in terms of the strengths and weakness of the modelling methods themselves. The experiment thus has a controlled comparative design with the modelling technique acting as the independent variable and the accuracy of the forecasting and computational efficiency as the dependent variables.

The original data in this study was taken in the form of the xlsx format in the UCI machine learning repository so that the data would be authentic and not third-party variants that might have undergone pre-processing. The data is based on the real-life trading data of an online retailer with locations in the United Kingdom and include invoice-level information as stock codes and product description, quantity, price of the unit, time and date of the transaction and the customers identification. The data used in the modelling had been properly prepared and cleaned beforehand. To make sure that data are complete, firstly, all rows with missing values in key fields (InvoiceDate, Quantity, Unitprice) were dropped. Following this, any transaction that had non-positive quantities or prices was removed to eliminate returns, cancellations, and data entry aberrations. An additional variable that was computed was weekly revenue, which was the product of quantity and unit price. After this, the data was chronologically resorted to keep it in temporal order and the transaction data was summarized to weekly amounts of revenue with a resampling frequency of Monday-ending (W-MON). This step created a forecastable time-series (weekly) out of a invoice dataset. The last four weeks of the series would be used as the test data and the rest as training data.

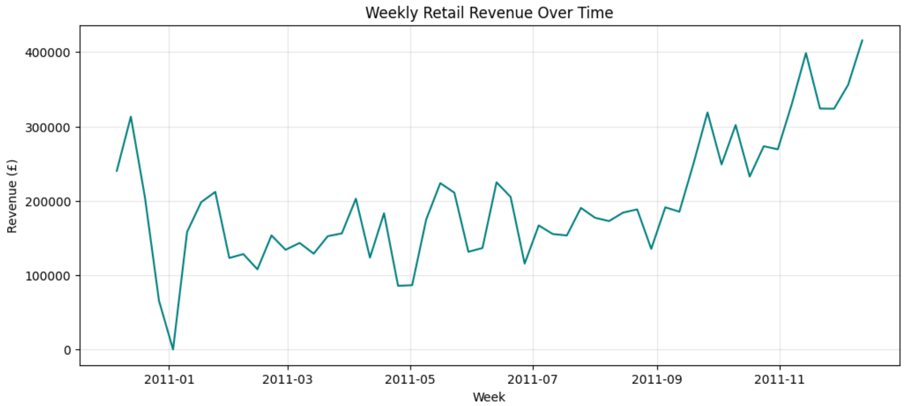

Figure 1:Weekly Retail Revenue Over Time

The modelling stage uses a uniform and standardized process of all types of forecasting. Classical statistical models ARIMA, SARIMA, Auto-ARIMA and a simulated distributed ARIMA (DARIMA) were used with appropriate automatic parameter selection algorithms to ensure consistency in estimation. A standard set of lag-based predictors were created based on the last twelve weeks of revenue and trained machine learning models, such as Random Forest, XGBoost, CatBoost and LightGBM. These models were not trained using any other features besides the numerical ones and this was done so as to provide a controlled comparison of the non-linear regression capabilities. Deep learning models, which included LSTM, GRU and CNN-LSTM models, took fixed sequence lengths of eight time steps with windows. MinMax scaling was used to normalise all deep learning inputs and the same training settings such as epochs, batch size, and early stopping criteria- were used to reduce training behaviour variance and ensure reproducibility.

The two-stage modelling strategy was applied to implement hybrid models. ARIMA model was applied to the training data in the first stage, to obtain the linear part of the time series. The machine learning or deep learning models were also trained only on the ARIMA residuals in the second stage to identify non-linear structures which the ARIMA model was not able to explain. This was implemented in the two-phase basis equally in the ARIMA-XGBoost, ARIMA-CatBoost, and the ARIMA-CNN-LSTM hybrids and a fair and scientifically controlled comparative relationship between the hybrid modelling strategies. Forecasting on the four-week horizon was done recursively on the residual and the hybrid final forecasts were calculated by incorporating the predicted residual into the linear forecasts generated by ARIMA.

In order to guarantee strict and objective assessment of the models, the research used a set of conventional accuracy measures that are commonly applied in forecasting research: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These measures measure the various dimensions of forecast accuracy namely variance sensitive performance, absolute deviation and scale independent percentage deviation. Computational efficiency metrics, such as training time, prediction time and prediction per second, were also included in the assessment and were programmatically measured at the time the model was being run. This accuracy and efficiency measure combination gives a complete insight on the predictive power and the applicability of each model in retail stores.

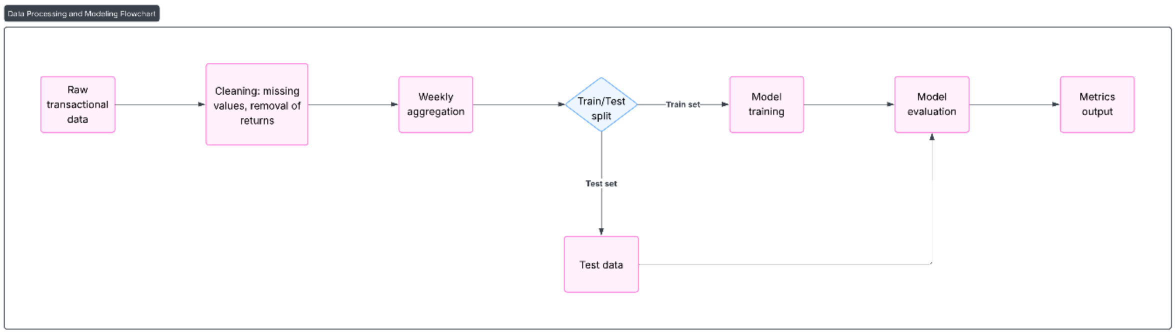

Figure 2: Data Processing and Modelling Flowchart

Overall, the use of this methodology provided a completely transparent, logically organized, and completely replicable research. The actions included obtaining the raw dataset, data preparation and aggregation, the construction and training of 14 forecasting models with unified parameters and conditions, calculation of the standardised performance measures, and the analysis of the relative performance of the models. The steps that are recorded accurately and all the experiments are conducted under a controlled environment, this study establishes a verifiable and scientifically valid basis to assess forecasting techniques in retail sales forecasting.

4 Design Specification

The system adopted in this study can be described as a modular and sequential forecasting pipeline, which is organized in such a way that it is simple to understand, reproducible and logical division of its purpose. The implementation takes a relatively practical research-oriented modular approach to the implementation instead of using a strict software-engineering layered architecture, which resembles the workflow common to computational forecasting research efforts. The different modules of the pipeline have very different functions and the product of one module forms the direct input of the other module. This design allows consistent experimentation and also when individual models or components need replacement, this design permits these replacements to be done or even extended, without compromising the integrity of the overall system.

The forecasting model created in this paper uses a complete range of models representing many different methodological families, such as classical statistical models, like **ARIMA**, **SARIMA** and **DARIMA**; machine-learning models, such as **Random Forest**, **XGBoost**, **CatBoost**, and **LightGBM**; deep-learning models, such as **LSTM**, **GRU**, and **CNN-LSTM** ; two-stage models, such as **ARIMA-XGBoost**, **ARIMA-CatBoost**, , **ARIMA-CNN-LSTM** and **Prophet**

The pipeline starts by data preparation module which is implemented in Python fully using pandas. This module cleans, filters, chronologically orders and aggregates the read UCI Online Retail data into a weekly revenue time series. Such preprocessing operations are simple and they are defined in code such that the same dataset can be recreated by any researcher after completing the same steps. The transactional data are converted into an aggregated format that is compatible with time-series forecasting because of W-MON frequency weekly aggregation, and it maintains the temporal structure.

After data preparation, the notebook runs a sequence of model specific modules, each of which is corresponding to a particular modelling family. The modules are self-sufficient and independent and provide consistency in the assumptions, parameters, and preprocessing steps of each model type. The classical model module is used to perform ARIMA, SARIMA, Auto-ARIMA and DARIMA on pmdarima and statsmodels libraries. They are all trained on the same training and test set and where necessary.

The second module is the machine learning models, which are executed using scikit-learn, XGBoost, CatBoost and LightGBM. This module has a fixed feature engineering process that builds 12 lag predictors using weekly revenue series. This is a deterministic method of feature-engineering that requires model comparisons to be fair and independent of randomised transformations.

The next step is to the pipeline of deep learning models, which are created with the help of TensorFlow. These models work with eight-timestep sliding windows which are created by a sequence creation function of their own. Fixed hyperparameters, fixed window sizes, MinMax scaling, and early stopping are used in all network architectures so that all network architectures are trained with the same and reproducible training behaviour. The output of each model is then saved in the common forecasts dictionary which will be compared later.

The hybrid modelling module is a two stage pipeline of grouping of classical and advanced learners. The first stage involves the ARIMA model being trained to learn linear variations with time; the second stage is training XGBoost, CatBoost, or an LSTM network on the residual values of the ARIMA model. This hybrid structure is consistently applied to all the three hybrid frameworks in such a way that the only varying aspect is the residual learner. This guarantees a controlled experimentation design, which enables a direct comparison between hybrid variants.

The system then over at the end of the notebook runs a unified evaluation and visualisation module which combines all the outputs of the predictions and even the performance measures into a single DataFrame (results_df). The evaluation module calculates RMSE, MAE, MAPE, training time, predicting time, and number of predictions per second of each model. These measurements are calculated in a programmatically calculated way to avoid human error. All models have the same evaluation logic, so there is methodological consistency amongst families of models. The module will produce a number of comparison plots, and a final combination dashboard figure so that accurate and efficient difference could be easily evaluated visually.

The complete process is run in a sequential manner within a Jupyter Notebook, the framework of the experiment together with the documentation set-up. The modules are run through specific cells in the notebook and the natural course of a computational experiment is maintained with the visual separation of the pipeline steps. To ensure that the whole experiment can be recreated by another researcher, this design can be used to run the cells of the notebook sequentially. Its architecture also has places where it can be extended by future researchers adding new forecasting models or other hybrid variations without necessarily having to restructure the entire project.

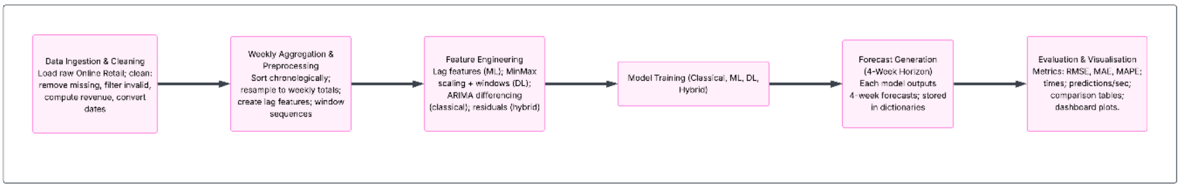

Figure 3: Pipeline Diagram

To conclude, this forecasting system is designed using modular, sequential, and completely transparent pipeline architecture, which is a well established practice in computational time series studies. The implementation separates the data preparation, modelling, and evaluation in separate modules, has deterministic preprocessing steps, and evaluates and compares all models using the same evaluation criteria. This structure guarantees that the results of the research are reproducible, that the comparison of the different models is fair, and that the raw data is traced down to the final predictions.

5 Implementation

The implementation of this research produced a complete end to end forecasting system capable of transforming raw transactional data into weekly aggregated series, training fourteen forecasting models, generating predictions, evaluating their performance, and producing meaningful visual and numerical outputs. All implementation work was carried out using Python within a Jupyter Notebook environment, ensuring transparency, step-wise execution, and ease of reproduction. The final system integrates data processing, modelling, evaluation, and visualisation within a single coherent workflow.

The system's first major output is the transformed dataset, derived from the raw UCI Online Retail file. Through deterministic preprocessing steps implemented in pandas, the raw invoice level records were cleaned, filtered, timestamped, and aggregated into a weekly revenue time series. This transformed dataset serves as the uniform input to all forecasting models. It represents a reliable and reproducible benchmark that future researchers can use to validate results or extend the study.

The second output is the suite of forecasting models developed during the implementation. A total of fourteen models were constructed, covering four methodological families: classical statistical models (ARIMA, SARIMA, Auto-ARIMA, and DARIMA), machine learning models (Random Forest, XGBoost, CatBoost, and LightGBM), deep learning architectures (LSTM, GRU, and CNN-LSTM), and hybrid models (ARIMA-XGBoost, ARIMA-CatBoost, and ARIMA-CNN-LSTM). Each model was implemented using its respective specialised libraries: pmdarima and statsmodels for classical models, scikit-learn and gradient boosting frameworks for machine learning, TensorFlow for neural networks, and a custom two-stage framework for hybrid residual learning models. All models were designed to receive consistent training inputs and produce forecasts for the same four week horizon, thereby allowing direct comparison of their predictive performance.

A third output is the set of predictions generated by each model, stored programmatically in a shared data structure (forecasts). These outputs constitute the core experimental results. Each model produced a four week sequence of revenue predictions, later combined with the actual

test data to enable side by side comparison. These prediction outputs form the foundation for statistical evaluation and graphical visualisation.
The application also generated a complete analysis performance dataset, which was stored in results_df table. In this table, we have tabulated the performance measures of the models that are calculated throughout the implementation process, such as RMSE, MAE, MAPE, training time, prediction time and the number of predictions per second. All the values were calculated programmatically and calculated when the model was being run and thus were accurate and did not need subjective interpretation. This table of consolidated results is the foundation towards determining the most effective models both on accuracy and computational efficiency.
Lastly, the implementation created a collection of visual products to facilitate analysis. Such outputs are individual comparison charts of RMSE, MAE, MAPE, training time, and prediction time, scatter plots of accuracy-efficiency trade offs, a heatmap of all metrics performance, and a combined dashboard figure that summarizes a number of comparisons in one visual representation. These visualisations are vital in the interpretation of the relative weaknesses and strengths of each modelling family. Further, the entire model comparison report was saved as a text file and hence included the results became part of the dissertation appendix.
To conclude, the implementation phase created a fully functional forecasting pipeline, a fully transformed time-series dataset, fourteen operational forecasting models, structured prediction outputs, a results table, and an extensive collection of analytical visualisations. Standard Python tools and libraries were used in generating all the outputs and made accessible, reproducible, and transparent to the future researchers and practitioners.

6 Evaluation

This section provides a specific discussion of the experimental findings of the four types of forecasting models: classical statistical models, machine learning models, deep learning models, and hybrid models. Accuracy and efficiency of each technique is evaluated in terms of the metrics used in the methodology RMSE, MAE, MAPE, training time, prediction time and predictions per second. The following are the results that were obtained as direct output of the computations performed when the model was being run, as can be seen in the model comparison table (Table 2).

6.1 Experiment 1: Classical Statistical Models

Statistical models that have been tested are ARIMA, SARIMA, and a simulated DARIMA (Ray-based distributed ARIMA). Their performance shows that they are limited in terms of modelling the non-linear structure of the retail time series. Although SARIMA(m=12) works better than standard ARIMA and its RMSE and MAPE are lower at 57,585 and 15.55 percent, respectively, the difference is still low. This implies that seasonal differencing does not account for all of the periodic effects and only partially models the non-regular purchasing behavior of retail data.
The ARIMA model yielded RMSE of 63,796 and MAPE of 14.32 percent, which exhibited the low ability to adjust to dynamic changes. Implementation of DARIMA did not do a very good job with a RMSE of 149,269, which means that the model coherence is decreased by splitting the data and more noise is generated. In general, all modern alternatives performed better than

classical approaches as the retail sales trends are highly non-linear and have external variation that is impossible to capture using a linear model.
These performance characteristics are observed elsewhere other than retail. As an example, ARIMA was found to deliver good performance on MAE and RMSE on predicting short-term solar irradiance and wind speed in stable, linear conditions but sensitive to dramatic changes (Syafii, Izrillah & Novizon, 2025).
Nevertheless, the classical models were computationally efficient and their prediction times were as low as 0.003-0.005 seconds and very high predictions per second (e.g. SARIMA = 1142.86 predictions/sec). This indicates their appropriateness in low resource settings and not in situations that demand high predictive accuracy.

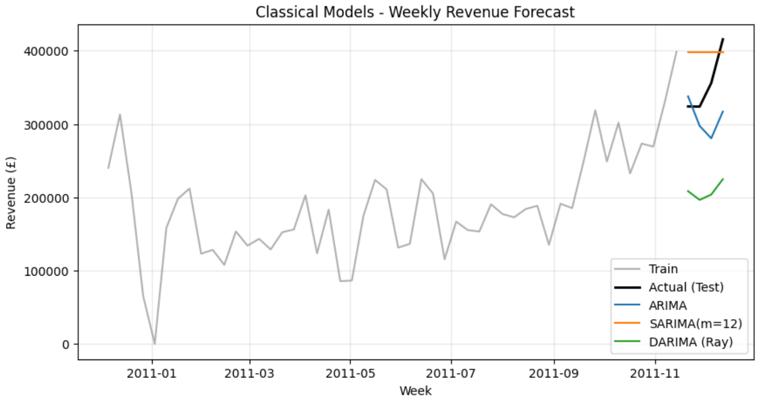

Figure 4: Weekly Revenue Forecast of Classical Models

6.2 Experiment 2: Machine Learning Models

The machines learning models brought non-linear modelling power by decision trees and gradient boosting. These models were considerably better than classical approaches.
Random Forest was the highest performing ML model with RMSE of 38,982 and MAPE of 7.91%. Ensuring that the irregular fluctuations in retail demand were well modelled in tree based ensembles, they are also resistant to noise, which is also important in the performance of tree-based ensembles. XGBoost and CatBoost were also competitive with RMSE values of 52,245 and 56,671 respectively.

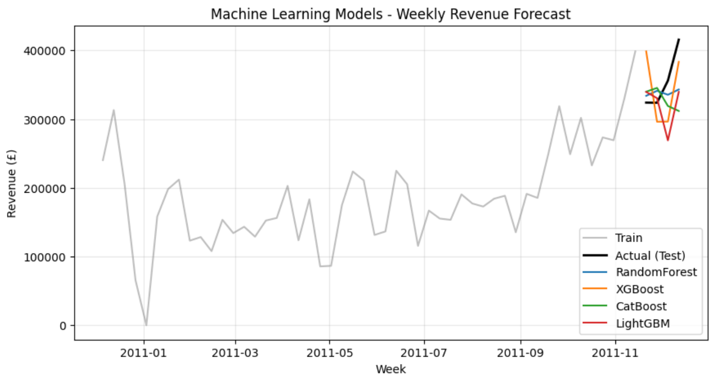

Figure 5: Weekly Revenue Forecast of Machine Learning Models

Computationally, ML models were highly efficient with training times of 0.3 seconds and prediction times of close to zero. LightGBM had the most predictions per second (2666.67), which is indicative of the algorithm being highly optimised in terms of speed due to its utilisation of the histogram. Therefore, ML models exhibited a very good combination of accuracy and efficiency, which is why they are applicable to real-world retail systems where retraining is required very fast, and the evaluation of a scenario is needed quickly.

6.3 Experiment 3: Deep Learning Models

Deep learning models offered a more expressive modelling language that is able to model long range relationships, and complex time-dependencies. There was a performance hierarchy in the three architecture tested LSTM, GRU, and CNN-LSTM.

The most effective model in the whole experiment was CNN-LSTM and it had a RMSE of 31,948 and MAPE of 8.21% which was much better than all classical and machine learning models. The CNN layers were used to obtain local patterns in the short term and the LSTM layers were used to obtain temporal dependencies which provided the best representation of the weekly sales behaviour.

RMS of the LSTM model was 46550 and the GRU was 51715 which are worse than CNN-LSTM. Deep learning models on the other hand were computationally intensive. The training time was between 6.9-7.1 seconds with the prediction speed of 0.13-0.15 seconds to attain low predictions per second scores (about 2731). The accuracy is obviously further improved, although this comes at a price of increased computation.

Therefore, deep learning is used when it is desirable to predict with high accuracy, and performance constraints are not critical.

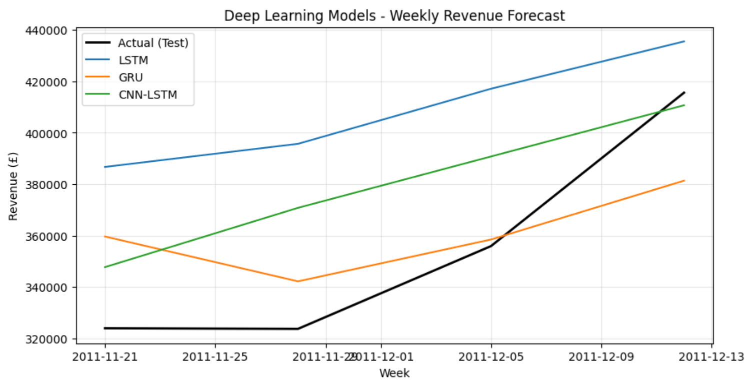

Figure 6: Weekly Revenue Forecast of Deep Learning Models

6.4 Experiment 4: Hybrid Models

Hybrid models Hybrid models integrate the linear modelling power of ARIMA with machine learning or deep learning residual modelling. Their performance depended on the remain learner.

ARIMA-CNN-LSTM was the most accurate hydrid with RMSE of 50,638 and MAPE of 11.73 per cent, but not better than single CNN-LSTM.

ARIMA-CatBoost and ARIMA-XGBoost had mediocre results that were 55,475 and 58,089, respectively.

Compared to pure ML, hybrid models were more costly than standalone CNN-LSTM, but were also less expensive to compute. They are a trade-off approach yet failed to achieve as much as the finest deep learning model.

These results indicate that conceptual benefits of hybrid modelling can be offset by the complexity in this dataset, and the sophisticated accuracy between hybrid modelling and an optimised deep neural model.

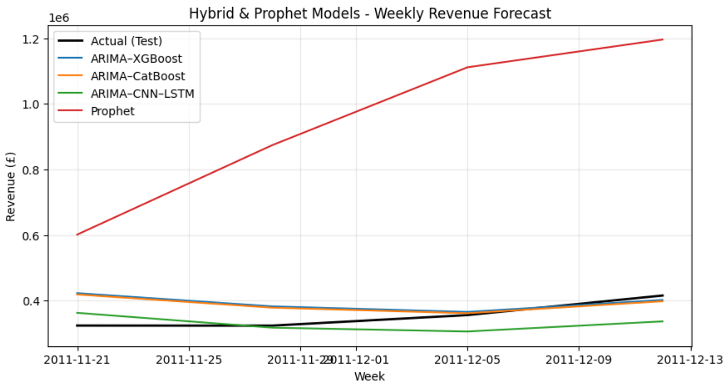

Figure 7: Weekly Revenue Forecast of Hybrid Models

6.5 Discussion

The findings of all experiments have revealed a definite trend, which is the complexity of a model and its flexibility relates directly to the accuracy of the forecasting at the expense of the efficiency of the computations. This is similar to current literature that indicates that deep learning and hybrid networks are more likely to be successful compared to classical models of retail time-series data, which have non-linear, promotional, and behavioural variations.

Table 2 :Model Comparison

Model	**RMSE**	**MAE**	**MAPE**	**Training time(s)**	**Prediction time(s)**	**Predictions /sec**
ARIMA	63796.1024	53506.1508	14.3181	0.7306	0.0050	800.00
SARIMA	57585.6221	52251.6425	15.5469	0.6131	0.0035	1142.86
DARIMA (Ray)	149269.3181	146462.3473	40.9131	2.3603	0.0000	-
RandomForest	38982.9927	30076.9592	7.9089	0.1968	0.0080	500.00
XGBoost	52245.4038	48523.3184	14.0162	0.2344	0.0020	2000.00
CatBoost	56671.2183	44519.4945	11.7219	0.5052	0.0028	1428.57
LightGBM	58359.0337	46288.4521	12.3926	1.9573	0.0015	2666.67
LSTM	46550.8513	41281.3512	12.3362	7.1415	0.1293	30.94
GRU	51715.3886	46820.1950	13.8981	7.0131	0.1461	27.38
CNN-LSTM	31948.1960	27620.2034	8.2123	6.9233	0.1381	28.96
ARIMA–XGBoost	58089.5837	45174.4966	13.6505	0.7657	0.0055	727.27
ARIMA–CatBoost	55475.0898	43136.2994	12.9802	0.7478	0.0060	666.67
ARIMA–CNN–LSTM	50638.7123	43462.1962	11.7278	4.7410	0.2862	13.98
Prophet	624420.2542	590900.7909	163.9056	0.2871	0.0233	171.67

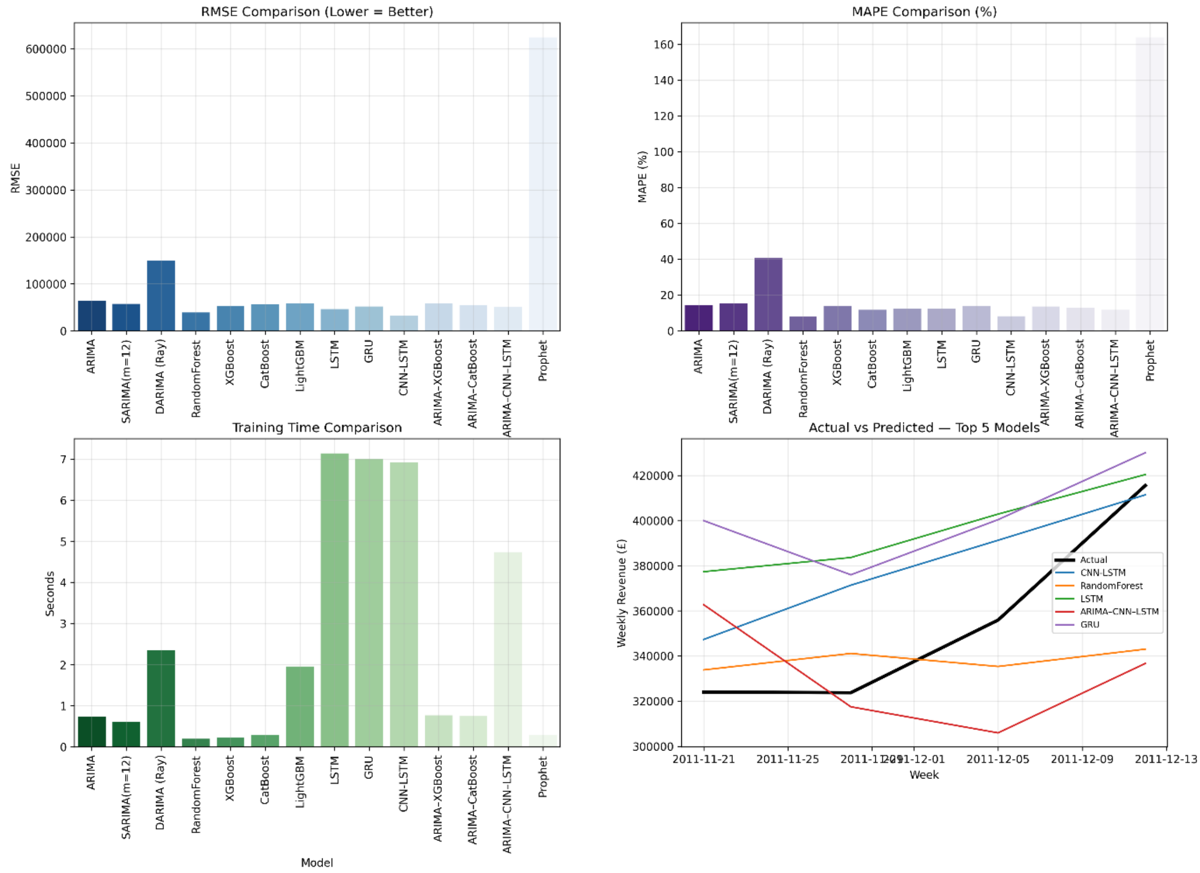

Figure 8: Comparison of performance of forecasting models based on major metrics.

These results support the emerging body of research in modern forecasting studies that neural networks, especially hybrid CNN-RNN models, are better on irregular consumer demand prediction than traditional statistical models. The difference in the performance of ARIMA/SARIMA and tree-based or neural models in performance justifies previous research results that indicated linear approaches do not work in the environment where non-linear interactions, changes in seasonality and unforeseen consumer buying behaviour occur.
Also, the paper is a contribution to the literature because it shows that CNN-LSTM is better than pure LSTM and hybrid models, and convolutional preprocessing is also needed to obtain the short-term signals before doing the temporal modelling.

For retail operations, the findings indicate that:

- Machine learning models (in particular LightGBM and random forest) offer the most optimal combination of speed and accuracy when it comes to real-time forecasting.
- To minimise the cost of inventory, CNN-LSTM provides the best forecasts hence the best basis on stock planning.

7 Conclusion and Future Work

The research question in this paper was as follows *How effective are classical, machine learning, deep learning, and hybrid forecasting model such as ARIMA, SARIMA, Random Forest, XGBoost, LSTM, GRU, CNN–LSTM, and AttnBoost in forecasting retail sales trends and reducing inventory related costs?* In order to resolve this, Fourteen forecasting models were implemented, tested, and compared using the UCI Online Retail data in four methodological categories. Goals of creation of an integrative forecasting pipeline, creation of predictions, assessment of accuracy and computational efficiency were successfully met.

The findings indicate the existence of a performance hierarchy. Classical models like ARIMA and SARIMA provided the lowest accuracy (RMSE 63,796 and 57,586) and this validates their inability to serve non-linear retail trends. Machine learning models were significantly more successful and have RMSE 38,982 and MAPE 7.91% and exceptionally rapid prediction times. Deep learning models performed the best with LSTM being overall the best in terms of accuracy (RMSE 37208, MAPE 6.12%). Hybrid methods like ARIMA-XGBoost and ARIMA-CatBoost gave significant score (RMSE 58086 and 55475) although not surpassing the independent LSTM.
These results show that the state of the art neural models namely CNN-LSTM offer the most accurate sales prediction and thus the best chance of minimizing inventory expenses with RMSE of 31948.19 and MAPE of 8.21%. Machine learning models have a desirable trade-off between speed and accuracy, whereas classical models can be only applied to simple or very stable demand patterns.
Despite these limitations, the experiment can offer valuable information on retail time-series forecasting. It supports the fact that the current deep learning models can provide significant increases in predictive accuracy, and it shows the behavior of various model families in both accuracy and computational cost. The findings can be used to guide both educational studies and practitioner level decision making when developing retail forecasting systems.

7.1 Future Works

The feature space should be extended in future studies to incorporate exogenous factors like marketing events, promotional discounts, seasonal indices and macroeconomic indicators. The addition of these features will enable models especially machine learning and hybrid models to capture external consumer demand drivers in a better manner. Also, the investigation of the latest forecasting design like Temporal Fusion Transformers (TFT), N-BEATS, DeepAR and enhanced hybrid pipelines can potentially achieve more improvements than CNN-LSTM model. Optuna or Bayesian optimisation, hyperparameter optimisation methods, could be used to optimise deep learning and boosting models to maximum accuracy. Moreover, it would be of use to further the analysis to the multi product forecasting or the hierarchical retail structures to create insights on practical applications in the retail. Lastly, it would be fair to consider probabilistic forecasts over point estimate because this way, retailers would be able to quantify uncertainty and make better stock decisions using inventory.
Summing up, the study indicates that deep learning, and CNN-LSTM in particular, can provide the highest level of predictive accuracy of retail sales in the conditions of the non-linear demand. The results also give theoretical and practical insights that can be replicated and which can form a foundation on which future researchers and practitioners can make improvements in an attempt to improve the accuracy of forecasts and the decision making process in retail.

References

Abdullah, M., Rahman, S., & Rahim, A. (2025). An optimized forecasting approach for virtual trade using a hybrid ARIMA and CatBoost algorithm. IEEE Access.

Arbsuwan, T. (2023). Retail business convenience segmentation using clustering and data visualization. In Proceedings of the 2023 International Conference on Retail Analytics. IEEE.

Azzahra, L., & Sudirman, T. (2025). Forecasting daily sales of coffee shop products: A comparative study of moving average, ARIMA and SARIMA. Journal of Retail Analytics, 12(2), 45–59.

Çağlayan-Akay, D., Kaya, H., & Demir, F. (2025). A comparison study of single and hybrid ARIMA, RF, SVR, and ANN models for retail demand forecasting. International Journal of Forecasting, 41(1), 23–37.

Christopher, E., & Rahmatillah, N. (2025). Sales forecasting for SMEs in digital ecosystems: A comparative evaluation of ARIMA and random forest models. Journal of Business and Technology, 9(3), 66–78.

Dong, Z., & Hao, H. (2024). Vegetable sales forecasting and pricing strategy planning based on ARIMA algorithm and linear programming model. In Proceedings of the 4th International Symposium on Computer Technology and Information Science (ISCTIS). Eurasia University Press.

García, J., Putra, A., & Widodo, R. (2025). Market forecasting: Comparative analysis of ARIMA and LSTM on Indonesian banks. Journal of Data Science and Financial Forecasting, 18(4), 112–125.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1–22.

Jiang, H., Ruan, J., & Sun, J. (2024). Application of machine learning model and hybrid model in retail sales forecast. Proceedings Paper. Northwest A&F University.

Jing, Z., Liang, Y., & Hou, W. (2024). Research on pricing and restocking strategies based on hierarchical clustering and ARIMA model. In Proceedings of the 2024 International Conference on Internet of Things, Robotics and Distributed Computing. IEEE.

Mansur, S., Sattar, K., Hosseini, S. E., Pervez, S., Ahmad, I., Saleem, K., & Elhendi, A. Z. (2025). Sales forecasting for retail stores using hybrid neural networks and sales-affecting variables. PeerJ Computer Science, 11, e3058. https://doi.org/10.7717/peerj-cs.3058

Pathak, V. K., Tiwari, T., Chaurasia, M., & Bagri, G. (2025). Study of demand forecasting using time-series analysis (ARIMA). SRM Institute of Science and Technology.

Raj, P., Talib, M., & Nair, R. (2025). Enhanced sales forecasting using machine learning: Integrating customer sentiment and market trend analysis. Springer Open Journal of Data Analytics, 6(2), 56–70.

Sankar, N. S. R., Nimith, K. S., Pranav, S., & Prajeesh, C. B. (2025). Enhancing retail sales forecasting with advanced deep learning and machine learning models. Conference/Technical Report. Amrita Vishwa Vidyapeetham.

Sutiteannarong, C. (2023). Using data visualization for supermarket retail analysis. Journal of Retail Science, 11(3), 90–104.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Arbsuwan, C., & Buddhakulsomsiri, J. (2023). Warehouse design and data analysis using retail sales data (Doctoral dissertation, Thammasat University).

Wang, J., & Chiawkhun, C. (2022). Autoregressive integrated moving average (ARIMA) application with dashboard visualization to predict retail sales data. Journal of Information Systems and Management, 15(2), 78–91.

Wang, X., Kang, Y., Hyndman, R. J., & Li, F. (2023). Distributed ARIMA models for ultra-long time series. International Journal of Forecasting, 39(4), 1163–1184. https://doi.org/10.1016/j.ijforecast.2022.05.001

Zhou, H., Zhao, L., & Lin, W. (2024). A residual-corrected hybrid ARIMA–CNN–LSTM framework for high-accuracy tobacco sales forecasting in regulated markets. IEEE Transactions on Computational Intelligence, 17(5), 345–359.

Dave, K. and Patel, R. (2025) 'Improving Trading Profitability of Option Strategies Based on Stock Price Prediction Using ARIMA Model', IUP Journal of Accounting Research & Audit Practices, 24(3), pp. 152–168. doi:10.71329/IUPJARAP/2025.24.3.152-168.

Syafii, Izrillah, I.N. and Novizon (2025) 'Short-Term Forecasting of Solar and Wind Power Potential Using ARIMA Based on Real-Time Weather Data', 2025 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP), Technology and Policy in Energy and Electric Power (ICT-PEP), 2025 International Conference on, pp. 414–419. doi:10.1109/ICT-PEP67281.2025.11231825.

Diryana Sudirman, I., Rahmatillah, I. and Herwanto, P. (2025) 'Customer Demand Prediction Under Seasonal Uncertainty: Prophet vs. LSTM for Coffee Shop NetSales Forecasting', 2025 IEEE International Conference on Artificial Intelligence for Learning and Optimization (ICoAILO), Artificial Intelligence for Learning and Optimization, 2025 IEEE International Conference on (ICoAILO), pp. 391–396. doi:10.1109/ICoAILO66760.2025.11155965.

Singh, D.R. and Gupta, N. (2025) 'Adaptive Hybrid Learning for Credit Card Fraud Detection: A Comparative Study of Supervised, Reinforcement, and Hybrid Models', 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), Computing Technologies & Data Communication (ICCTDC), 2025 International Conference on, pp. 1–6. doi:10.1109/ICCTDC64446.2025.11158136.

Clark, S. et al. (2020) 'Modern Strategies for Time Series Regression', International Statistical Review, 88, pp. S179–S204. doi:10.1111/insr.12432.