Department of Computer Science and Engineering

Consultancy Project Sample

PROJECT COMPLETION REPORT

Title of the Project:
Demand- Driven Forecasting in Supply Chain Management Using Ensemble Learning Approach.

Principal Investigator

Dr. Lipika Goel,
Associate Professor
Department of Computer Science & Engineering
Gokaraju Rangaraju Institute of Engineering and Technology,

Bachupally, Kukatapally, Hyderabad,
Telangana, INDIA, PIN Code: 500 090.
Phone No. +91 9811296119
Email: lipika1670@grietcollege.com

Dr. K.Madhavi,
Professor & Head
Department of Computer Science & Engineering
Gokaraju Rangaraju Institute of Engineering and Technology,

Bachupally, Kukatapally, Hyderabad,
Telangana, INDIA, PIN Code: 500 090.
Phone No. +91 9849364116
Email: hod_cse@griet.ac.in

TECHNICAL INFORMATION

Demand forecasting being a critical business requirement to support a range of business functions, including logistics, production, and finance. Demand forecasting is the process of making estimations about future customer demand over a defined period using historical data. While traditional methods have been extensively used in demand forecasting due to the recent boost in the AI world, companies have started researching the possibility of using machine learning algorithms in place of traditional approach. In this study, supply chain demand is forecasted with different algorithms applied on a dataset comprising the sales data with a time of four to five years and their results are compared.

The existing system heavily relies on manual processes, making it time consuming and prone to human errors. This manual handling of data increases the risk of inaccuracies and reduces the overall efficiency of the forecasting process. Additionally, the current models in use may face challenges in adapting to changing market conditions or unexpected events. As a result, the forecasts generated by the system may be less reliable and fail to provide actionable insights during dynamic and uncertain market scenarios. Furthermore, the system's focus on long-range forecasting may lead to a neglect of the accuracy of short-term forecasts. This can impact operational decision making, as timely and precise short-term forecasts are essential for effective inventory management, resource allocation, and meeting customer demands.

Businesses employ various methods to forecast demand, utilizing data and analytics over specific time periods. These approaches can be categorized into short-term and long-term forecasting.The proposed framework utilizes application of machine learning algorithms for long term and short-term demand forecasting. It plots the trends and the predictions for the given time is made.

Short-term demand forecasting is typically conducted for a period of less than 12 months. It aims to analyse demand for a shorter duration to support day-to-day operations. In contrast, long-term demand forecasting is carried out for periods exceeding a year. It assists in Identifying and planning for seasonality, annual patterns, production capacity, and expansion plans over an extended timeframe. This type of forecasting drives long-term business strategy.

WORK PLAN AND DELIVERABLES

In order to achieve the aim and objectives of the proposed system, the following is the work was the stated plan and deliverables.

Literature review

Field study

Study of the Machine Learning approaches

Data acquisition and Understanding

Development of the framework for demand prediction

Providing prediction reports.

In order to understand how machine learning can be used in demand forecasting, we researched several forecasting techniques that have been developed to solve diverse forecasting problems. Since each forecasting problem has its own purpose, challenge, and application, finding the right methodology is crucial to have an accurate performance. We explored the traditional, statistical approaches and the algorithmic approaches that have been developed in recent years. Later, we further studied some of the real industry application of long-range forecasting examples in diverse cases focusing on methodologies, variables, and performances.

Demand Forecasting Methods Forecasting methods fall into two categories: subjective and objective. Subjective forecasting methods can be further subdivided into judgmental and experimental. These forecasting methods are used when historical data is not available, as in the case of a firm launching a new product. Objective forecasting methods are either causal or timeseries approaches. In causal methods, there is a relationship between a set of variables, which is then used to formulate a forecast. Time series methods are used when there is a pattern in the data.

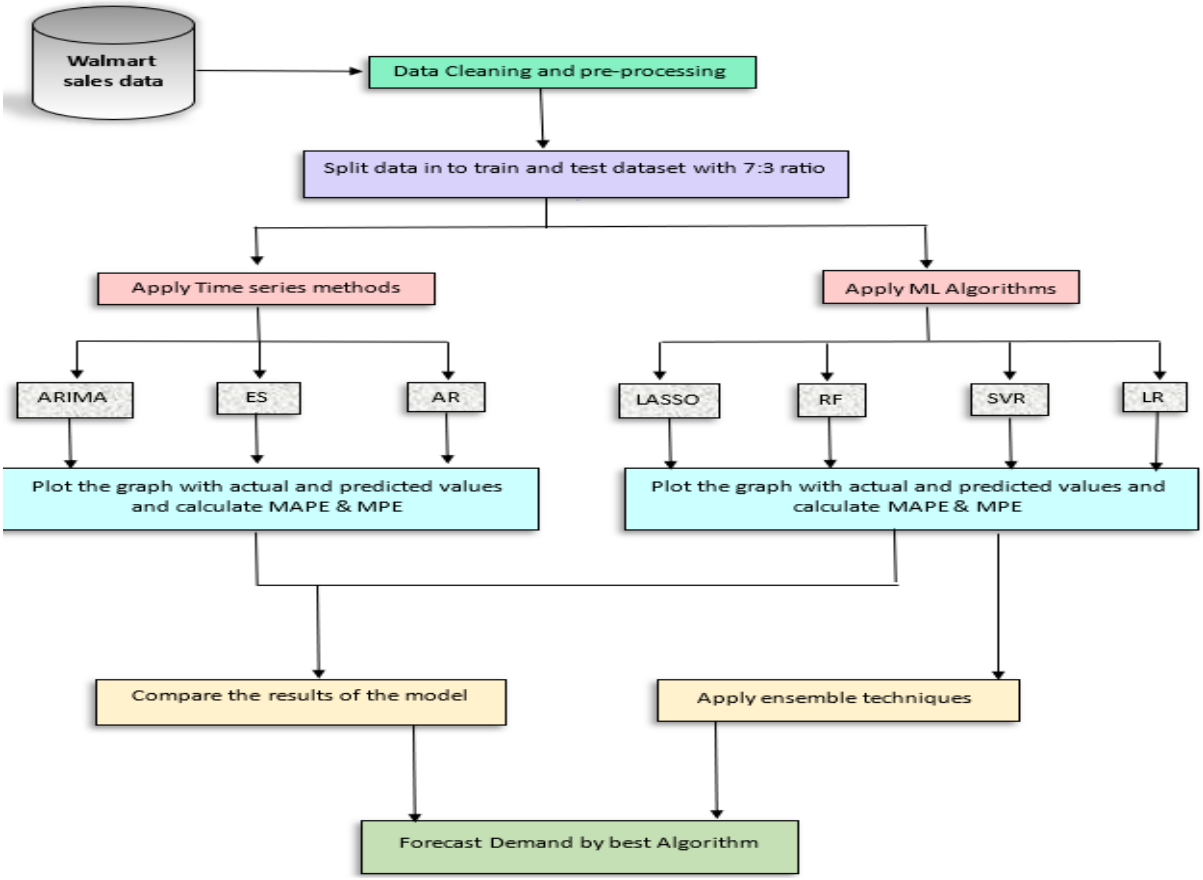
Through the literature review, we found that no forecast method is always right for every circumstance. Among multiple methodologies available for demand forecasting, the selected forecasting model should fit well with the task considering data type, data availability, data horizon, number of factors that affect the result and character of the trend itself.

First, we reviewed the data the sponsor company provided. Since the methodology used in a forecasting model depends on the type of variables, dataset, and forecasting purpose, we decided to test the feasibility of machine learning methodology for one specific category of product. This product had the most accessible data available for our purposes. Next, we studied general machine learning methodologies, ensemble approaches and the time series methods for both short term and long-term forecasting. After investigating some of the prevalent machine learning methodologies used to improve forecasting practices, such as support vector machine (SVM), Linear Regression, random forest, Lasso regression, we identified which methodologies to test on the sponsor company’s data.

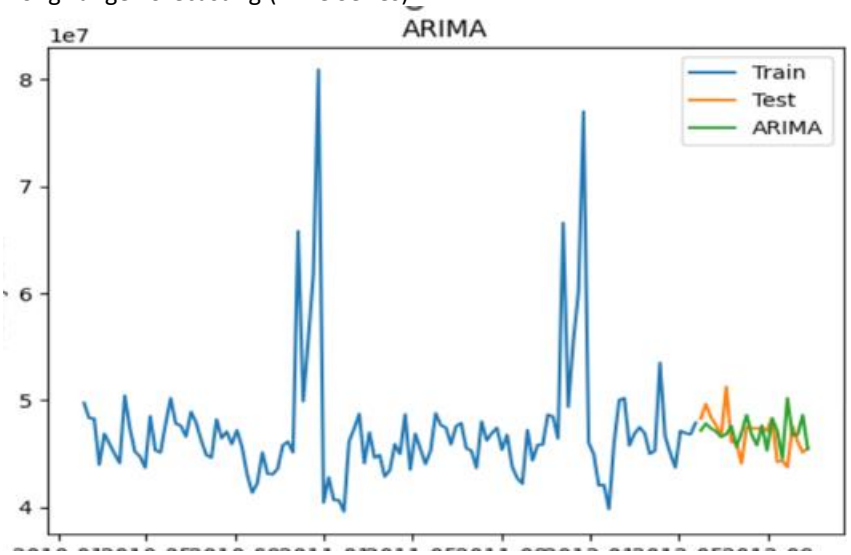
The data acquisition step included the gathering of the available data and looking through a statistical summary for understanding of the data. The data was cleaned analyzed for gaps and anomalies, checked for relevance, and

restored. Once the data was cleaned, generated, and checked for relevance, it was structured into a comprehensive form.

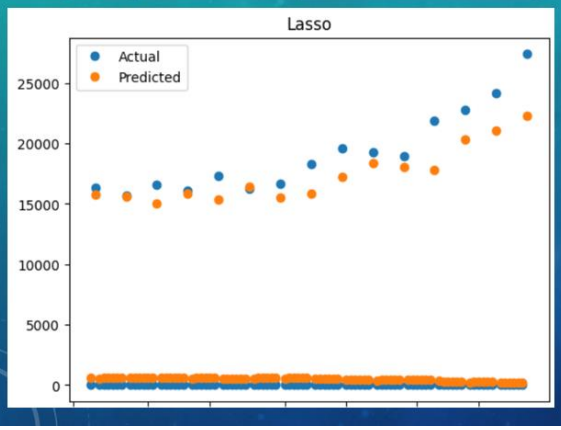
After compiling the data, variables were standardized and only the most relevant features were selected through the feature selection method. Direct modeling only uses historical data to forecast future time. Direct modeling approach was tested using time series algorithms and four different machine learning methodologies: random forest (RAN), support vector machine (SVM), Lasso Regression and linear regression (LR). The ensemble strategy, which combines the strengths of various machine learning algorithms, also increases forecast accuracy is also included.

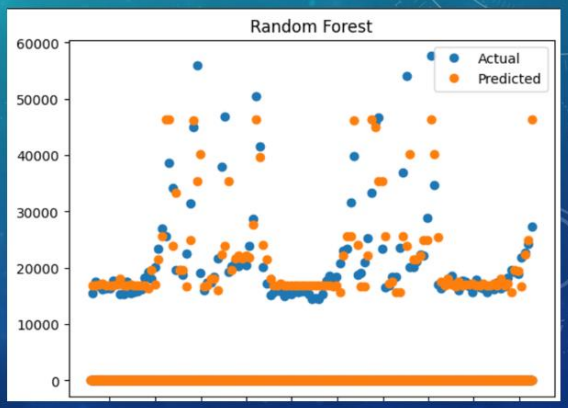


Long-range Forecasting:(Time Series)

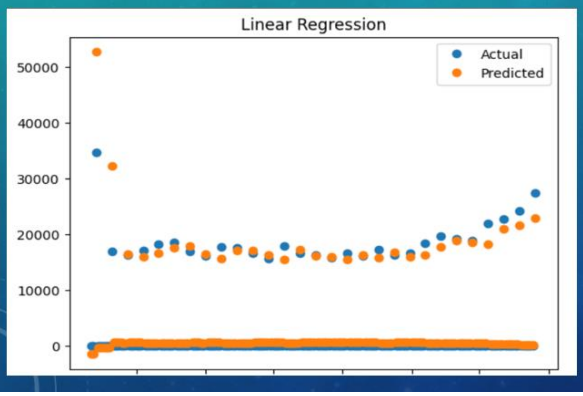


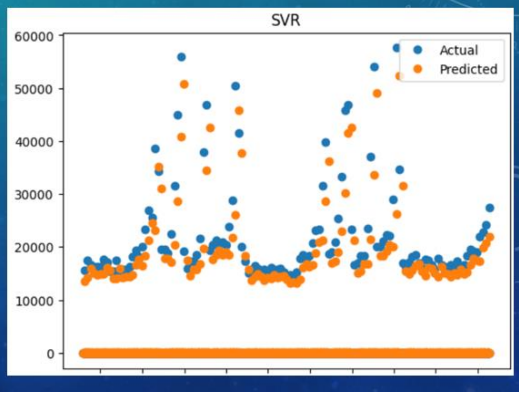
Long-range Forecasting:(Machine Learning)



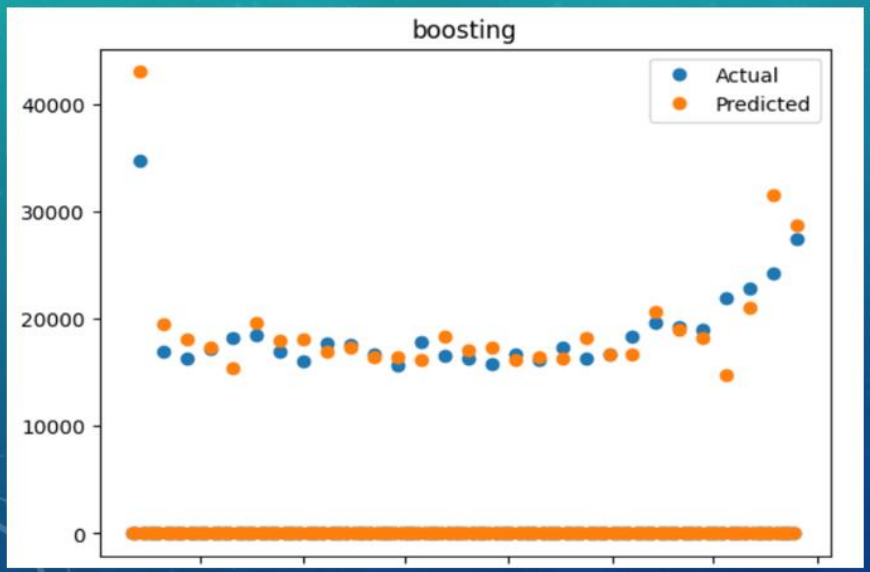


Long-range Forecasting:(Machine Learning)

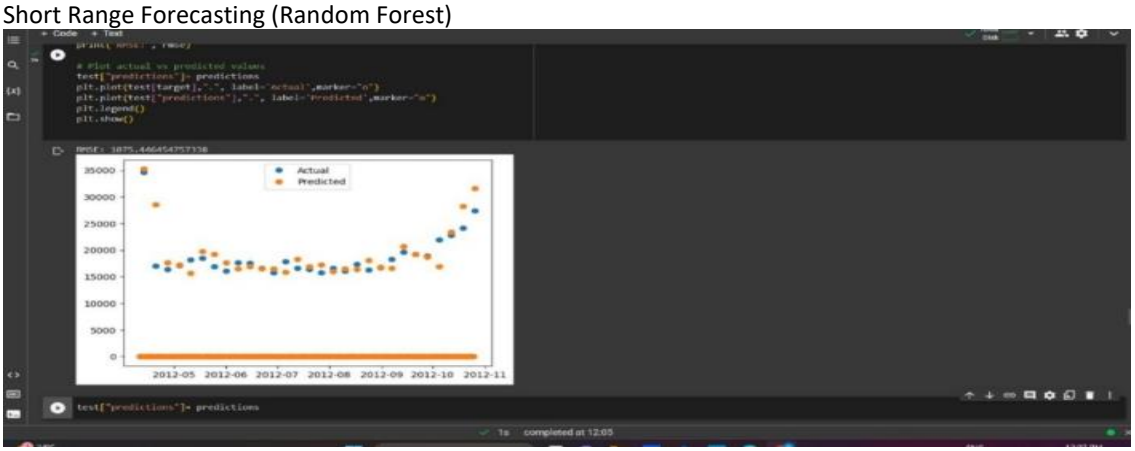




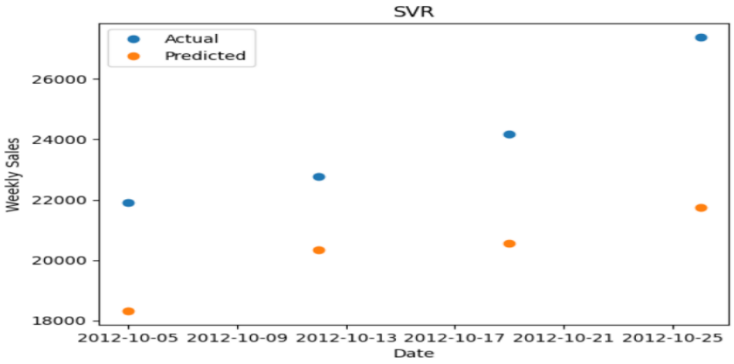
Ensemble:



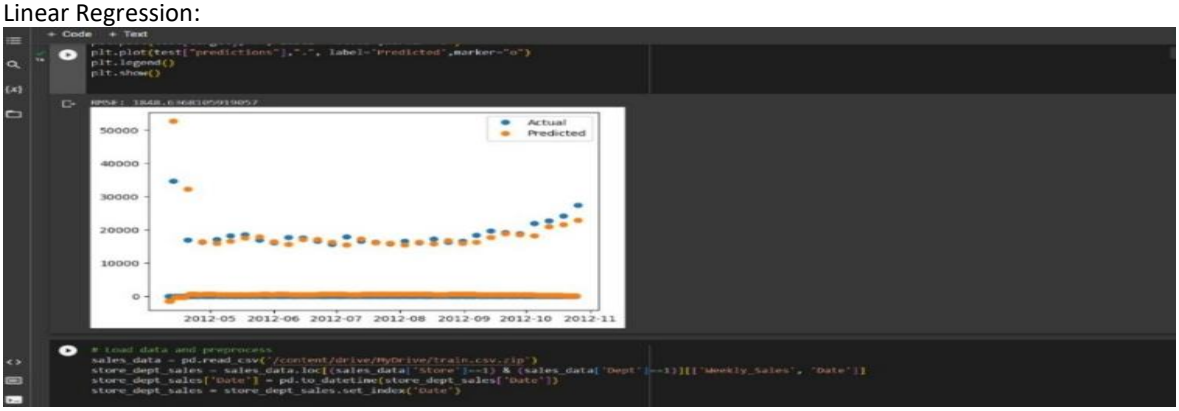
Short Range Forecasting (Random Forest)



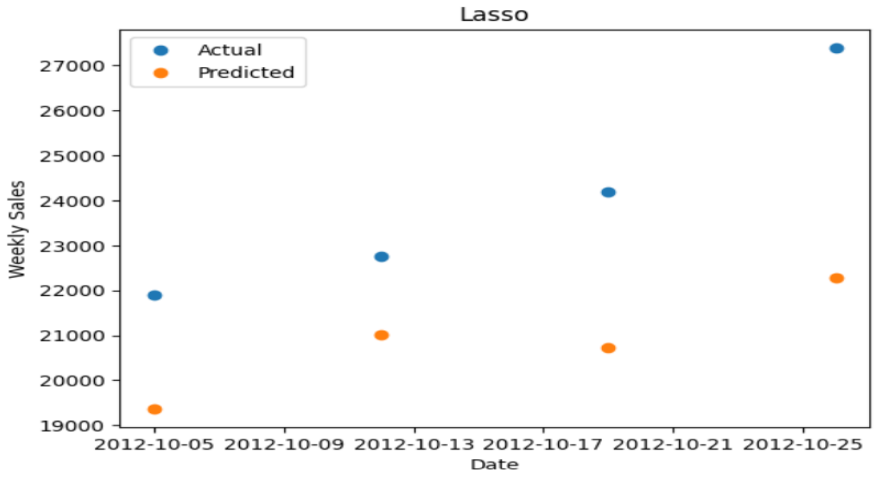
SVR:



Linear Regression:



Lasso Regression:



ENSEMBLE APPROACH:

We have used Random Forest and XGBoost as ensemble approach for predictions. Hyperparameters of Random Forest and XGBoost were optimized using only training data. 5- fold cross-validation method was applied to evaluate the training accuracy of the model. With random grid search the parameters were tuned. The model was evaluated via the mean average percentage error. We applied boosted trees to predict future demand. The below table tabulates the MAPE values for both short term and long term demand forecasting using different approaches of time series and machine learning.

Table: MAPE values of Short term and long term demand forecasting

Category	Technique	Models	MAPE
Short Term Demand Forecasting	Time Series	ARIMA	9.28%
	State of art Machine Learning Algorithms	Linear Regression	10.23%
		LASSO Regression	13.05%
		ANN	24.70%
		SVR	15.66%
	Ensemble	Random Forest	10.46%
		XGBoost	15.76%
Long Term Demand Forecasting	Time Series	ARIMA	33.16%
	State of art Machine Learning Algorithms	Linear Regression	10.77%
		LASSO Regression	8.64%
		ANN	20.78%
		SVR	30.22%
	Ensemble	Random Forest	10.03%
		XGBoost	7.98%

Time Series ARIMA performed best in short term forecasting with MAPE of 9.28%, followed by LR 10.23% and RF with 10.46%. MAPE. In general, long-term prediction on group level is much more accurate with ensemble approach with MAPE as low as 7.98% followed by Random Forest with 10.03% MAPE.

Developing and providing accurate demand forecasting models helps intermediary companies to remain competitive and maintain or increase their profit margin by optimizing inventory management and making informed, data-driven decisions. Future research will investigate and compare the potential of autoregression and filtering methods as well as deep learning methods for time series prediction, such as Recurrent Neuronal Networks (RNNs) and Long Short-Term Memories (LSTMs) to improve the demand forecasting.

Department of Computer Science and Engineering

PROJECT PROPOSAL FOR ADINTEL PVT LTD

1. **Title of the Project:**

AN OCR BASED INVOICE AUTOMATION SYSTEM FOR SUPPLY CHAIN MANAGEMENT

Principal Investigator
Dr. Ashlin Deepa R N,
Associate Professor, Department of Computer Science and Engineering,
Gokaraju Rangaraju Institute of Engineering and Technology,
Bachupally, Kukatapally, Hyderabad,
Telangana, INDIA, PIN Code: 500 090.
Phone No. +91 8500282924
Email: ashlin887@grietcollege.com

Co-Principal Investigator
Dr.P.Chandra Sekhar Reddy
Professor, Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology,
Bachupally, Kukatapally, Hyderabad,
Telangana, INDIA, PIN Code: 500 090.
Phone No. +91 99894 41542
Email: pchandra1369@grietcollege.com

TECHNICAL INFORMATION

1. SCOPE OF THE PROJECT

Optical Character Recognition (OCR) creates a digital version of a printed, typed, or handwritten document that computers can read. It avoids the need to manually type or enter text. Scanning the document and running OCR software quickly turns the file into something which can be edited.

In asset-intensive industries, critical warehouse processes are still completely paper-based, inefficient and error prone. The first step to solving a poor goods receipt process is to replace the existing paper-based process with a mobile application. OCR algorithms can automatically digitize these documents, extract the information, and pipe them into a database for storage, alleviating the need for large, expensive, and even error-prone manual entry teams. OCR technology has grown and matured over the years with breakthroughs in Artificial Intelligence.

Invoice OCR refers to the process of extracting relevant data from scanned or PDF invoices and converting it into a machine readable format that is both editable and searchable. The process of digitizing an invoice can be broken down into 3 steps:

1. Converting the physical document to a digital variant - this could be done through invoice scanning i.e., clicking an image through a camera
2. Information Extraction - this can be done by

- Humans - manually done by reviewers who will analyse the invoice for errors, read the text in it and enter it into a software for storage and future retrieval.
- Machines - Optical Character Recognition (invoice OCR) - recognizing the text and numbers present in the documents.
- Information Extraction - once the Process of OCR is complete it's important to identify which piece of text corresponds to which extracted field. If a field is the total, subtotal, date of invoice, vendor etc.

3. Data dump - once the information has been extracted it needs to be stored in a retrievable format like

- A database
- An excel sheet
- An ERP system.

1. To automate processes

With deep learning and OCR, you can automatically take these invoice images, extract tables and text from them, extract the values of different fields, make error corrections, check if the products match your approvable inventory and finally process the claim if everything checks out. This is a massive leap from what the insurance industry has traditionally done, but it can prove very beneficial nevertheless.

2. To increase efficiency

By digitizing invoices, several processes can be made a lot faster and smoother. Take for example a retail store chain that deals with a few regular vendors for commodities and process payments at the end of every month. This store can save a lot of time by automating the process of invoice management. Vendors just have to upload the bills on an app or a website and they can get instant feedback on if the images are of good resolution if the image is of the entire invoice if the image is fake or was digitally manipulated, etc saving a lot of time.

3. To reduce costs

The same retail stores' franchise saves a lot of money by automating invoice digitization using PDF OCR and deep learning. An invoice which has to pass through the hands of three reviewers so there are no errors reduces to one. The number of invoices processed by a computer is several times faster than what a human could do. The time includes checking if the invoice is a fraud, if it has all the information, if all the information is correct, entering all of the data manually into a spreadsheet or a database, running calculations and finally processing the payment.

4. For better storage

In the case of disputes, the vendor can reach the app and look through all the invoices he/she uploaded and the post-processing results of each invoice, explaining the commodities, their quantities, the costs of each, the taxes and the discounts. The company, having automated the process of entering this data into a database, can also now retrieve this information anytime.

2. SCIENTIFIC RATIONALE FOR THE PROPOSED WORK

Nowadays, we are witnessing the dramatic process of digital transformation in which digital technologies are changing how businesses connect and create value for their customers [1]. Digital transformation is gaining an important role in the production and supply chain applications. As the name suggests, OCR is used to recognize and extract the characters from the image [2]. These characters can be anything, from letters in different languages to numbers. In this research, a software solution capable of recognizing serial numbers from wine bottles is presented.

The languages provide greater challenges specifically to classifiers, and also to the other components of OCR systems. Google is committed to making its services available in as many languages as possible [3], so we are also interested in adapting the Tesseract Open Source OCR Engine [4, 5]. Tesseract began as a PhD research project [6,7] in HP Labs, Bristol, and gained momentum as a possible software and/or hardware add-on for HP's line of flatbed scanners. After a joint project between HP Labs Bristol, and HP's scanner division in Colorado, Tesseract had a significant lead in accuracy over the commercial engines, but did not become a product.

3. OBJECTIVES OF THE RESEARCH

- To support all processing stages in an OCR system with a single format
- To represent layout information in terms of a generative typesetting model
- To encapsulate existing OCR engine output formats
- To represent and associate information from different stages of OCR processing

4. EXPECTED OUTCOME AND SIGNIFICANCE OF THE PROPOSED PRODUCT

These large organizations employ data entry teams whose sole purpose is to take these physical documents, manually re-type the information, and then save it into the system. Optical Character Recognition algorithms can automatically digitize these documents, extract the information, and pipe them into a database for storage, alleviating the need for large, expensive, and even error-prone manual entry teams. OCR is a technology that enables computers to read text from images, and it can be used to automate invoice-processing tasks. This project aims to develop an OCR for extracting data from invoices and other accounting documents, and converting them into searchable and editable text.

5. WORK PLAN AND DELIVERABLES

In order to achieve the aim and objectives of the proposed system, the following is the work plan and deliverables.

1. Literature review

2. Field study
3. Invoice management
4. Identifying best Invoice processing systems
5. Implementation
6. Provide progress report to the client

6.METHODOLOGY

Invoice OCR refers to the process of extracting relevant data from scanned or PDF invoices and converting it into a machine readable format that is both editable and searchable.

Digitizing an Invoice

The process of digitizing an invoice can be broken down into 3 steps:

a) Converting the physical document to a digital variant - this could be done through
   - invoice scanning
   - clicking an image through a camera

b) Information Extraction - this can be done by
   - Humans - manually done by reviewers who will analyse the invoice for errors, read the text in it and enter it into a software for storage and future retrieval.
   - Machines -
     - Optical Character Recognition (invoice OCR) - recognizing the text and numbers present in the documents.
     - Information Extraction - once the Process of OCR is complete it's important to identify which piece of text corresponds to which extracted field. If a field is the total, subtotal, date of invoice, vendor etc.

c) Data dump - once the information has been extracted it needs to be stored in a retrievable format like
   - A database
   - An excel sheet
   - An ERP system.

7. TIME SCHEDULE OF ACTIVITIES AND MILESTONES

Activity	0	3	6	9	12	15	18	21	24
Grant Received									
Literature survey									
Implementation									
Testing									

Results						[redacted]

PUBLICATIONS OF PI:

- Ashlin Deepa R.N, "Document Text Analysis and Recognition of Handwritten Telugu Scripts", Proceedings of 4th International Conference on Cybernetics, Cognition and Machine Learning Applications, ICCCMLA 2022, 2022, pp. 462–466
- Ashlin Deepa RN and Rajeswara Rao R,"An Eigen characters method for Recognition of Handwritten Tamil Character Recognition", Proceedings of the First International Conference on Intelligent Computing and Communication, Advances in Intelligent Systems and Computing 458, Springer 2017, Page(s): 495-505.
- Ashlin Deepa, R.N. and Rajeswara Rao R, (2016). "An efficient offline Tamil handwritten character recognition system using zernike moments and diagonal-based features", International Journal of Applied Engineering Research, ISSN 0973-4562, Vol.11, No. 4, Page(s): 2607-2610.
- Ashlin Deepa R N. and Rajeswara Rao R, (2017) 'A modified GA classifier for offline Tamil handwritten character recognition', International Journal of Applied Pattern Recognition, Inderscience Publishers, ISSN 2049-8888, Vol. 4, No. 1, Page(s): 89–105.
- Ashlin Deepa R N and Rajeswara Rao R, (2019) "Classification of Handwritten Tamil Characters using Variable Length Puzzle Pieces", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Vol. 8, No. 12, October 2019, Page(s): 4797-4801.
- Ashlin Deepa, R.N. and Rajeswara Rao, R. (2020), A novel Nearest Interest Point classifier for offline Tamil Handwritten character recognition, Pattern Analysis and Applications, SPRINGER, ISSN: 1433-7541, Vol. 23, No.1, Page(s):199-212.

Patents:

National Patent: Title of the invention: MACHINE LEARNING MODEL FOR PREDICTING SEVERITY PROGNOSIS IN PATIENTS INFECTED WITH COVID-19.

PUBLICATIONS OF CO-PI:

1. P Chandra Sekhar Reddy, "Energy Efficient Data Retrieval in Wireless Sensor Networks for Disaster Monitoring Applications", International Conference on Sustainable Energy and Future Electric Transportation (SEFET), 2021.
2. P Chandra Sekhar Reddy, "Predict the Performance Analysis of Supervised Learning Techniques Using Heart Disease Database", 2nd International Conference for Emerging Technology (INCET), 1-7, 2021.
3. P Chandra Sekhar Reddy, "Improve the integrity of data using hashing algorithms", International Journal of Innovative Technology and Exploring Engineering, 2019.
4. P Chandra Sekhar Reddy, "Detection of Malicious uniform Resource Locator", International Journal of Recent Technology and Engineering (IJRTE), 2019.
5. P Chandra Sekhar Reddy, "Schedule-based Optimized Node Recharging Model (SONRM) For Increasing Longevity of Wireless Rechargeable Sensor Network (WRSN)", 2019