Natural Resources for Human Health
eISSN: 2583-1194

Original Research

**Received** 16 May 2026
**Revised** 20 June 2026
**Accepted** 22 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 383–415

**KEYWORDS**:
Biomass Supply Chain, Bioethanol, Fuzzy Data Envelopment Analysis, Machine Learning, Supply Chain Optimization, Deep Learning

# A Hybrid ML–FDEA Framework for Efficient and Uncertainty-Aware Biomass Supply Chain Optimization

Dipali I. Ghadage[1], Dr. Jagannath Nalavade[2], Dr. Ramachandra Pujeri[3]

[1] Research Scholar, Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India 412201,
Email id: dipali.ghadage@ritindia.edu, Orcid Id: 0009-0007-4108-9143
[2]Professor, Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India 412201
Email id: jen20074u@gmail.com, Orcid Id:0000-0002-1645-0095
[3]Pro-VC, MIT Art Design and Technology University, Pune 412201, India.
Email: sriramu.vp@mituniversity.edu.in, Orcid ID: 0000-0003-2588-6728

**ABSTRACT:** This paper introduces a new hybrid model, combining both Machine Learning (ML) and Fuzzy Data Envelopment Analysis (FDEA), to optimize the biomass supply chain management in bioethanol production systems. The study concentrates on the effective transportation of excess sugarcane bagasse at the sugar mills to bioethanol factories in Western Maharashtra region. To simulate the dynamics of a real-world supply chain, a detailed multi-year dataset (2017-2025) was built based on production, transportation, and cost-related parameters gained through official reports, and then preprocessed and engineered features according to the systematic standards. The FDEA model proposed assesses the efficiency of plant-mill transportation routes by considering various inputs and outputs under uncertainty, with the efficiency scores being between 0 and 1. Through the analysis, it is found that there is a high variability in route performance as the values of efficiency lie between 0.072 and 0.465, illustrating the effect that transportation distance and cost have on the efficiency of the supply chain. In order to have better scalability and predictability, machine learning models were trained to predict FDEA efficiency scores. Ensemble learning methods such as AdaBoost, Random Forest and Decision Tree were found to perform strongly on predicting with R2 of 0.950, 0.943 and 0.947 respectively and custom deep learning as implemented in PyTorch and TensorFlow were found to outperform with R2 of 0.975 and 0.978 respectively. The findings confirm that the hybrid ML-FDEA model is an effective method to incorporate efficiency assessment, uncertainty modelling and predictive analytics, such that the biomass supply chain optimization framework offers a strong decision support system. The framework has a better accuracy, scalability and viable applicability than the current methods. The research makes a contribution to sustainable bioenergy systems, whereby biomass logistics can be optimized cost-effectively and by data to optimize biomass logistics in the face of real-world uncertainties.

## 1. INTRODUCTION

The shift in the world to the sustainable system of energy has put more emphasis on renewable energy sources and more so biofuels that are made out of biomass. Lignocellulosic biomass is turn into bioethanol, which is viewed as a promising alternative to fossil fuels, as it has the potential to mitigate greenhouse gas emissions, increase energy security, and help with the implementation of a circular bioeconomy [1,2]. Sugarcane bagasse, a large agricultural by-product produced in the sugar production process, has received much interest as a prospective raw material in second-generation bio-ethanol production, among other biomass feedstocks. The combination of its high content of cellulose and hemicellulose makes it appropriate in biochemical and thermochemical conversion process in current biorefineries [3]. Big amounts of bagasse are produced in the main sugar-producing areas, especially in the developing economies. Though some part of this biomass is used internally to do cogeneration in sugar mills, a significant proportion of it is excess and it is underutilized. The effective exploitation of this excess bagasse has the potential to play a major role in sustainable fuel production and waste reduction measures. The effectiveness of these biomass-to-biofuel systems, however, is greatly influenced by the layout and functionality of the related supply chain that encompasses various steps such as gathering, transportation, storage, and distribution [4]. Because of the reliance on spatially-located sources of feedstock, fluctuating rates of production, and numerous logistical limitations, the biomass supply chain is an inherently complex system. Specifically, bio-ethanol production schemes need continuous and consistent biomass supply, or agricultural residues like bagasse are produced periodically. This supply-demand mismatch compels the need to develop effective and responsive supply chain management (SCM) systems that can maximize the distribution of resources and reduce the operating expenses [3].

The efficient management of bagasse supply chains is a major issue although interest in biomass-based energy systems has been growing. Among the main problems is the seasonal and geographically dispersed character of bagasse availability that causes inconsistencies to the supply and logistical inefficiency. Large amounts of bagasse are generated in a short period of time during peak crushing sessions, but bio-ethanol plants need to be fed on a year-round basis. This time discrepancy makes it more difficult to manage the inventory, storage planning, and transportation scheduling. A second dire issue is that low-density biomass materials have a high transportation cost. Bagasse is voluminous and of low energy density and hence transporting it is economically costly especially when covering a long distance. The cost of transportation depends on various factors including distance, fluctuations in fuel prices, vehicle capacity and route conditions all of which create a lot of variability in the system. Moreover, the uncertainty of the parameters fuel cost, traffic conditions, biomass yield and others also complicate the process of decision-making in supply chain optimization. In real-world scenarios, many of these parameters are not deterministic but uncertain and dynamic in nature. Such uncertainties are not frequently reflected in the traditional SCM models, and hence the solutions to them are suboptimal or unrealistic. Thus, there is a dire necessity in sophisticated modeling frameworks that will be capable of multi-objective optimization, uncertainty management, and predictive decision-making in biomass supply chains.

An overview of the current literature indicates that biofuel supply chain design has been addressed by a number of optimization methods, such as mathematical programming, stochastic optimization and multi-objective models. An example is recent works combining mathematical modeling and artificial neural networks to improve bioethanol supply chain planning and resilience [3]. On the same note, hybrid optimization models that integrate predictive analytics and efficiency assessment methods have also been designed to enhance site selection and logistics planning in biofuel systems [5]. Nevertheless, current methods have a number of major limitations. One, the majority of supply chain optimization models are deterministic i.e. they assume accurate input parameters of cost, distance, and demand. This is not realistic in biomass supply chains where uncertainty is intrinsic owing to the environmental, operational, and economical factors. Despite the fact that some studies involve stochastic or probabilistic modeling, it may take a lot of computation and does not adequately reflect the vagueness of real-world data. Second, Data Envelopment Analysis (DEA) has been extensively applied to efficiency analysis in supply chains but traditional DEA models operate on crisp data, and they do not effectively address uncertainty. The latest progress has also brought about fuzzy extensions of DEA, which have not been fully applied in the optimization of the biomass supply chain. Third, machine learning (ML) methods have been used in more and more ways in the prediction and decision-making of energy systems, but their combination with efficiency analysis tools, like DEA, remains insufficiently studied. Current literature usually applies ML or optimization models in isolation, as opposed to building a single model that incorporates predictive and efficiency-related decision-making. Hence, there is a massive gap in the research to come up with an integrated hybrid model that incorporates machine learning with fuzzy DEA to enable uncertainty-sensitive optimization of the supply chain in biomass-based bio-ethanol systems.

According to the research gaps identified, the main questions of this research are as follows:

- To come up with a data-based structure of analyzing and optimizing the performance of biomass supply chain with real-world production and logistics data.

- The aim of the research is to model and assess the efficiency of mill-to-plant transportation routes when Fuzzy Data Envelopment Analysis is used under uncertainties.
- To incorporate machine learning algorithms to forecast the efficiency of the supply chain and to be able to support dynamic decision-making.
- To develop an efficient supply distribution plan that will reduce transportation cost and still deliver biomass at all times to the bio-ethanol plants.

The research contributes to the sphere of optimization of biomass supply chains and sustainable energy systems in several ways:

- **Hybrid ML-FDEA Framework:** This paper suggests a new hybrid framework of machine learning and fuzzy data envelopment analysis to improve supply chain decision-making when faced with uncertainty.
- **Monthly DMU-Based Modeling:** The idea of month-wise decision-making units (DMUs) is proposed to be able to perform efficiency analysis of transportation routes more accurately and in a context-specific way.
- **Uncertainty Modeling With Fuzzy Logic:** Triangular fuzzy numbers are used to model uncertainty in transportation distance and fuel cost, enhancing the strength and realism of the model.
- **Bioethanol SCM Decision Support System:** The created system is a practical decision support system which is going to help identify optimal mill-to-plant routes, enhance the use of resources, and lower the overall cost of the logistics.

This paper takes the present state-of-the-art in optimization of biomass supply chains, combining predictive analytics, efficiency evaluation, and uncertainty modeling into a single framework. The proposed solution is a scalable and practical method of sustainable bio-ethanol production systems and also valuable to policy makers, industry stakeholders and researchers engaged in renewable energy supply chains.

## 2. LITERATURE REVIEW

The role of optimizing biomass supply chains in achieving sustainable production of bioenergy has greatly been researched. In biomass supply chain, various stages are interrelated in that, the collection of feedstock, preprocessing, transportation, storage and conversion involve various stages leading to overall system efficiency and cost [5]. The complexity and multi-dimensionality of the design of an efficient supply chain is caused by the spatial and temporal variability of biomass resources. A number of researches have been directed towards the development of optimization models to supply chains in biomass based on mathematical programming techniques. Zahraee et al., [6] have emphasized that the mixed-integer linear programming (MILP) is among the most widely used tools owing to computational tractability and feasibility to solve large-scaled optimization problems. Nonetheless, such models tend to be deterministic and do not reflect the real-life uncertainties. Recent developments have brought forth strong and evidence-based optimization models to deal with uncertainties in biomass supply chains. To enhance decision making under varying biomass availability and cost conditions, Huang et al., (2025) have come up with a data-driven robust optimization framework with uncertainty sets. Likewise, Alinejad et al., (2024) have also mentioned the significance of circular economy concepts and intelligent approaches to biomass resources management and the necessity of integrated and flexible supply chain models. There is also interest in hybrid solutions that use predictive analytics and optimization. Wang et al., [8] developed a two-step model that incorporated artificial neural networks with optimization algorithms to enhance biofuel supply chains by selecting optimal sites and optimizing operations. Abandansari et al., [9] also evidenced the usefulness of fuzzy inference systems and decision-making systems in studying the interactions of complex biomass supply chains and in determining the best strategies. Nevertheless, in spite of such developments, the current research remains limited in terms of the possibility to combine the notions of uncertainty modeling, efficiency analysis, and predictive analytics into a single framework. This underscores the necessity to have hybrid models that can be able to deal with several aspects of biomass supply chain optimization at once.

Data Envelopment Analysis (DEA) is a method of non-parametric efficiency analysis that has been widely applied in the supply chain management. DEA is a method used to assess the relative effectiveness of decision-making units (DMUs) based on the comparisons between a large number of inputs and outputs, which is why it is especially applicable to logistics and transportation systems [10] .DEA has found extensive application in supply chain applications to measure the efficiency of operations, performance of suppliers, and effectiveness of logistics networks. Guerrero et al., [11] used DEA to determine the efficiency of transportation systems and the best strategies of resource allocation. In the same manner, Hahn et al., [12] have also shown how DEA can be used to benchmark the performance of a supply chain in various industries. Recent research expanded the uses of DEA to energy and biofuel supply chains. Hong, J. [13] applied DEA and optimization models to assess the efficiency of site selection in biofuel supply chains, and demonstrated a higher accuracy of decision-making. According to Razmi et al., [14], DEA-based models can be used to assess the biomass logistics performance, but they have limited predictive potential and are typically not integrated with machine learning methods. Nonetheless, conventional DEA models are constrained by the fact that they require the use of accurate input-output data, which is not realistic in the practical supply chain systems. These shortcomings require the formulation of improved DEA models that are able to manage uncertainty and dynamism.

In order to eliminate the shortcomings of traditional DEA, fuzzy DEA extensions have been created to enable uncertainty and imprecision in input-output data. Fuzzy DEA (FDEA) permits fuzzy numbers to represent uncertainties, which give more realistic efficiency tests of complex systems. Singh et al., [15] came up with a Hybrid fuzzy multi-objective DEA model of assessing DMUs in uncertain environment and showed it to be more robust than the traditional DEA models. Bolos et al., [16] also contributed to the development of fuzzy DEA techniques by combining different fuzzy models and graphical analysis software which facilitated a more detailed evaluation of efficiency in uncertain settings. Fuzzy DEA has been applied in the energy and supply chain to assess performance under uncertainty where the cost, demand, and operational conditions are uncertain. Habibi et al., [17] highlighted that the inclusion of uncertainty in supply chain models is critical in enhancing decision reliability and resilience of the system. In a similar fashion, Gital et al., [18] emphasized uncertainty in biomass supply, cost and logistics as having significant effects on supply chain performance and should be mitigated by adopting advanced modeling methods. Although it has its benefits, the use of fuzzy DEA in the optimization of biomass supply chain is still in its infants. The majority of the studies either concentrate on fuzzy modelling or optimization as a single entity and never combine the predictive analytics and machine learning to make better decisions.

Machine learning (ML) has become a potent predictive modeling and decision-support tool in the supply chain management. ML methods facilitate data-based predictions, pattern identification, and optimization and can, therefore, be used on complex and changing systems like biomass supply chains. The review of the ML application in biomass supply chains by Peng et al., [19] indicated that it was effective in predicting biomass availability, logistics optimization and enhancing the efficiency of the supply chain. Peng and Sadaghiani [20] showed how ML techniques could be used in the evaluation of biomass, storage and transportation with an emphasis on enhancing operational performance. In energy systems and optimization of biomass, deep learning methods also have been implemented. Hai et al., [21] used deep learning models to optimize bioenergy systems and proved to be more accurate in predicting system performance. Zhao et al., [22] pointed out the increased role of ML in forest biomass supply chains especially in making predictions and risk management. The concept of integrating ML and optimization models has been examined in recent studies to help improve the decision-making process. Han and Zhang, [23] used neural networks and optimization methods to enhance supply chain design and efficiency. Nevertheless, Zhang et al., [24] observed that the combination of ML and efficiency evaluation tools like DEA remains a sparse area of research and is a critical research gap. Moreover, Khedr et al., [25] stated that the resilience of the supply chain may be enhanced with the help of ML models that allow predicting and making decisions in real-time and under uncertainty. Zhao et al., (2022) also emphasized the opportunities of AI-based strategies in improving the performance of the bioenergy systems with predictive analytics and intelligent optimization.

Based on the above analysis of literature, a couple of gaps in the existing body of research can be identified as critical. To begin with, the majority of biomass supply chain optimization models are highly deterministic, in the sense that they assume fixed values of the parameters that include transportation cost, distance and availability of biomass. But biomass supply chains in the real world are uncertain because of seasonality, unpredictable fuel costs, and disruptions in operations. Current deterministic models do not reflect this uncertainty and provide less reliable and practical solutions. Secondly, Data Envelopment Analysis (DEA) has become a popular efficiency measurement tool in the supply chain systems because it can deal with numerous inputs and outputs. Nevertheless, conventional DEA models are founded on specific numerical data and cannot be applied to uncertainty or ambiguity in real-life situations. This restricts their use in the optimization of biomass supply chains where there is a high variability in data. Third, Fuzzy Data Envelopment Analysis (FDEA) has been established to deal with uncertainty by using fuzzy logic in the DEA models. Although FDEA offers a more realistic method of assessing efficiency in the face of uncertainty, it has not been widely used in biomass supply chain systems. The majority of the existing studies are based on the theoretical development of fuzzy DEA without the incorporation of this into a practical supply chain optimization model. In addition, machine learning (ML) methods have also proved to be effective in predicting, recognising patterns and providing decision support in supply chain systems. ML models have been effectively applied in predicting the availability of biomass, optimizing logistics, and improving system performance. These models are however applied as stand-alone models and are not combined with efficiency evaluation methods like DEA or FDEA. Consequently, the existing ML-based methods do not have a benchmarking and selection of the most efficient supply chain routes. Also, the literature tends to treat optimization, uncertainty modeling, and prediction as discrete issues, instead of creating a comprehensive framework that encompasses all three elements. This disjointed process constrains the efficiency of supply chain decision-making and lessens the viability of available models. As such, a significant and evident research gap exists on the development of a hybrid framework that merges machine learning with fuzzy Data Envelopment Analysis to the optimization of biomass supply chain under uncertainty. This framework can be used to simultaneously solve prediction, efficiency analysis, and uncertainty modeling to achieve a more robust, realistic and practical solution to bio-ethanol supply chain management. The given gap is the basis of the current research that seeks to design a comprehensive ML-FDEA-based supply chain management framework that will be able to maximize the biomass transportation in the face of real-world uncertainties.

The literature has indicated a number of significant gaps in the optimization of biomass supply chain. The majority of the models available are deterministic and are based on set parameters like transportation cost, distance and availability of biomass, which do not reflect the real-world uncertainties caused by seasonal variations and volatile fuel prices. Despite the popularity of Data Envelopment Analysis (DEA) as an efficiency evaluation method, it makes several assumptions about accurate data and does not represent uncertainty well. Fuzzy DEA (FDEA) helps overcome this limitation but is mostly confined to the theoretical research without the practical implementation into the supply chain systems. In the meantime, machine learning (ML) tools have good predictive power, but are generally used alone, without combination with efficiency analysis tools. Therefore, existing methods consider optimization, uncertainty and prediction independently. This brings to the fore the necessity of an integrated hybrid ML-FDEA model of sound and uncertainty-ready biomass supply chain optimization.

**Table 1:** Literature Review

| Study | Method | Uncertainty Handling | ML Used | Limitation |
|---|---|---|---|---|
| Abandansari et al., (2023) | Fuzzy Inference System | ✓ | × | Does not integrate DEA efficiency analysis |
| Guerrero et al., (2022) | DEA-based SCM Evaluation | × | × | Deterministic DEA, no uncertainty modeling |
| Hahn et al., (2021) | DEA Benchmarking | × | × | Static efficiency evaluation only |
| Hong (2023) | DEA + Optimization | × | × | Lacks uncertainty handling and prediction capability |
| Razmi et al., (2025) | DEA for Biomass SCM | × | × | No integration with ML, limited decision support |
| Singh et al., (2022) | Fuzzy Multi-objective DEA | ✓ | × | Not applied to biomass supply chains |
| Bolos et al., (2025) | Advanced Fuzzy DEA | ✓ | × | Lacks integration with real-world SCM systems |
| Habibi et al., (2023) | Uncertainty-based SCM Models | ✓ | × | No efficiency evaluation mechanism |
| Gital et al., (2024) | Biomass SCM under Uncertainty | ✓ | × | No predictive modeling or ML integration |

| | | | | |
|---|---|---|---|---|
| Peng et al., (2023) | ML-based Forecasting | × | × | No optimization or efficiency evaluation |
| Peng & Sadaghiani (2023) | ML for Biomass Logistics | × | ✓ | No integration with DEA or SCM optimization |
| Hai et al., (2023) | Deep Learning Optimization | × | ✓ | Limited interpretability and no efficiency scoring |
| Zhao et al., (2024) | ML for Biomass Forecasting | × | ✓ | No integration with optimization models |
| Han & Zhang (2023) | ML + Optimization | × | ✓ | Does not incorporate uncertainty or DEA |
| Zhang et al., (2024) | AI-based SCM Review | × | ✓ | Identifies gap but does not propose hybrid model |
| Khedr et al., (2023) | AI-driven SCM | × | ✓ | No efficiency benchmarking framework |

## 3. METHODOLOGY

The purpose of this study is to come up with an effective Supply Chain Management (SCM) model to streamline the transportation of excess sugarcane bagasse at sugar mills to the bioethanol production facilities in the Western Maharashtra region. The suggested system will aim at reducing transportation distance and costs at the same time increasing the effective supply of surplus bagasse to bioethanol plants and hence enhance the overall efficiency of logistics and the use of resources. In order to accomplish this goal, the hybrid methodology framework that combines Fuzzy Data Envelopment Analysis (FDEA) and Machine Learning (ML) is used. To analyze the relative efficiency of mill-to-plant transportation routes, FDEA model is employed to analyze numerous input and output parameters under uncertain circumstances. Simultaneously, machine learning models are created to forecast FDEA efficiency scores using significant transportation, cost and production-relevant variables. This combined solution increases the ability to determine the best supply chain paths and offers a data-driven decision support system to the biomass transportation planning. The methodology overall includes systematic data retrieval, preprocessing, estimation of excess bagasse, transportation costs modeling, efficiency estimation using the FDEA, and predictive analysis using machine learning, which constitute a full-scale framework of supply chain optimization.

### 3.1 Data Extraction

The dataset to be used in this research was found in the annual reports of the Sugar Commissionerate of Maharashtra State, Pune, [27] during the years 2017 to 2025. These reports are in Portable Document Format (PDF) and they have detailed information that is associated with sugar mills consisting of sugarcane crushing, sugar production, and generation of bagasse. But, as the data was presented in an unstructured PDF form, as opposed to a machine-readable structured dataset, a systematically structured process of data extraction was needed to transform the data into a structured tabular format that could be analyzed and used in supply chain modelling. The extracted dataset contains various important variables; among them, the name of the district, name of the sugar mill, mill code, type of ownership, daily crushing capacity, start and end dates of the crushing season, total sugarcane crushed, sugar production, sugar recovery percentage, duration of crushing, and percentage of bagasse produced. These variables are the basis of future supply chain modelling, efficiency analysis and predictive analysis.

Extraction of the data was done in Python, through PDF parsing libraries namely Camelot and PDF Plumber which are readily available tools of extracting structured and semi-structured data out of PDF files. Camelot was used to process well-defined tabular data with clear table structures that were identifiable, and PDF Plumber was used to process irregular tables or those where data was stored in text format. With these tools, data of various PDF reports were systematically downloaded and turned into structured pandas DataFrames, which allowed to handle, transform, and preprocess data effectively. This move worked well in converting unstructured PDF data into a structured data that could be used to perform subsequent analytical and modelling operations. Most PDF extraction libraries written in Python, like Camelot and PDF Plumber, are popular to extract tabular data in PDF documents and convert them into structured datasets to analyze and model the data. AI-Based Table

Extraction, 2025) by Cyril Sadovsky, however, was initially in Marathi, which means that it had to be translated into English, so it could be compatible with the processes of analysis and modelling. To this end, the Google Translator library (Deep Translator) of Python was used to translate the column headers, as well as the textual data values, into English [29]. This translation process guaranteed consistency and readability and usability of the dataset in the future data preprocessing, analysis and machine learning based modelling.

### 3.2 Data Cleaning

After extraction and translation of the dataset to English, a thorough data cleaning exercise was carried out to verify the quality of the data, consistency and reliability. Data cleaning is an essential phase of data-driven research, especially when the dataset will be used to perform more complex data processing, such as estimating surplus bagasse, modelling transportation costs, Fuzzy Data Envelopment Analysis (FDEA) and predictions through machine learning. The extracted data had a few problems such as missing values, inconsistency in translations, inconsistent categorical entries and wrong data types that needed to be preprocessed systematically before further analysis. First, inconsistencies of the translation process were resolved. It was noted that some categorical variables like the names of districts, types of ownership and identifiers of sugar mills were either not translated correctly or they were not consistently presented. To address this, text standardization and manual correction were done to make all the categorical fields consistent. This is a necessary step because the inconsistency of categorical variables may result in improper grouping, biased statistical analysis, and decreased model accuracy in machine learning applications.

All the variables of the datasets were then transformed into the correct data types in order to be able to compute and analyze the data correctly. Numerical variables, such as crushing capacity per day, crushed sugarcane, sugar production, percentage sugar recovery and bagasse production were coded into numeric forms as well as temporal variables, such as date of crushing season start and date of crushing season end were coded into standard date formats. String or categorical data type was used to encode categorical variables such as district, type of ownership, name of plant, and type of vehicle. Correct data type alignment guarantees that mathematical operations, statistical assessments and model training are conducted correctly. The other important aspect of the cleaning process was to handle missing data. The key variables related to production were found to have missing values in the variables of sugarcane crushed, sugar recovery percentage, and sugar production. With these parameters, it is crucial to the proper estimation of bagasse and the analysis of efficiency, all the records with the missing values in key fields were eliminated to preserve the data integrity. Also, some numerical fields had zero values which meant missing or unrecorded data. The values were turned into null values to avoid distortions of cost modelling, efficiency calculations, and machine learning predictions.

Crushing season start and end date were also used as temporal feature engineering. These date variables were broken down into day, month and year parts, so that the duration of crushing could be calculated as well as it was possible to analyze seasons. These types of insights in time are especially relevant in the biomass supply chain modelling since bagasse supply is not only seasonal but also directly affects transportation planning and logistics optimization. All yearly datasets were processed individually and then combined into a single multi-year dataset to allow long-term trend analysis and effective modelling of the yearly datasets. After the merging, missing values in the numerical and temporal variables were filled with the values of the means of the districts. This method has been taken with the assumption that sugar mills in the same district will have similar operational characters, production capacity and bagasse production trends hence giving real and context specific estimations. The last processed and cleaned dataset was the basis of the next steps, such as surplus bagasse estimation, transportation costs modelling, estimation of efficiency based on the FDEA model, and machine learning model development to optimize the supply chain. Table 2 provides a systematic overview of data cleaning methods used in this study and summarizes them, giving an idea of how preprocessing was done and what its purpose was.

**Table 2** Data Cleaning Techniques Used in the Study

| Cleaning Technique | Description | References |
|---|---|---|
| Manual translation and text standardization | Incorrect translations and inconsistent categorical values were corrected and standardized | [30] |

| | | |
|---|---|---|
| Data type conversion | Numerical, categorical, and date variables converted into appropriate data types | [30] |
| Handling missing values | Rows with missing values in critical production variables were removed | [31] |
| Zero value treatment | Zero values representing missing data were converted into Nan. | [32] |
| Date feature extraction | Day, month, and year extracted from crushing season dates for seasonal analysis | [30] |
| Merging multi-year dataset | Yearly datasets (2017–2025) merged into a single dataset | [30] |
| Missing value imputation | Missing values filled using district-wise mean | [31] |

### 3.3 Dataset Creation

The main aim of this work is to design a Supply Chain Management system using Machine Learning and Fuzzy Data Envelopment Analysis (FDEA) to optimize transportation of surplus sugarcane bagasse on the basis of finding efficient mill-to-plant paths. The data available on the Sugar Commissionerate of Maharashtra consisted of a few variables regarding sugar production and sugar crushing like the district, type of ownership, daily capacity of the crushing, sugarcane crushed, sugar output and sugar recovery percentage. Nevertheless, in order to carry out the optimization of the supply chain and to apply the FDEA model, other variables pertaining to bagasse production, transportation, and the demand of the plant were needed. Thus, a number of new variables were inferred and added to the dataset, such as bagasse production, excess bagasse, the demand of the plant, mill and plant identifications, the distance between mills and plants, the fuel cost, transportation cost, transportation cost per ton of bagasse, vehicle capacity, and a monthly surplus bagasse dataset. These variables played a critical role in modelling every mill-to-plant route as a Decision-Making Unit (DMU) in the FDEA model where efficiency is measured by minimizing transportation cost, fuel consumption and distance and maximizing bagasse supplies. Methodologies and variables in existing research studies on biomass-to-biofuel supply chain optimization informed the choice and development of these features, which guaranteed reliability and applicability of the developed model.

**Percentage of Bagasse Production:** The percentage of bagasse production was determined by an empirical correlation between sugar recovery and bagasse production in the sugar production process. Increased sugar recovery leads to decreased bagasse production and the other way round. The percentage of bagasse production was determined with the help of the equation number 1 (Eq. (1)).

$$\text{Bagasse }(\%) = 27.41 - 0.3 \times (\text{Sugar Recovery}) \quad \textbf{(1)}$$

The value that was constant was adjusted with the real industry data to enhance precision. This equation presents a useful estimate of the percentage of bagasse production to help estimate the total bagasse and excess supply in the supply chain model.

**Bagasse Production (Tonnes):** The total bagasse production was estimated using the quantity of sugarcane crushed since bagasse is a direct by-product of the crushing process. The following equation (Eq. (2)) was used to estimate the bagasse production in tonnes.

$$\text{Bagasse Production} = \frac{\text{Sugarcane Crushed (Tonnes)} \times \text{Bagasse Production }\%}{(100)} \quad \textbf{(2)}$$

**Calculation of Surplus Bagasse:** According to FAO studies, much of the bagasse is utilized in sugar mills as fuel, but with increased efficiency, a section of the total bagasse can be utilized as surplus to be used outside the mills like the production of bioethanol [33]. Report by FAO Forestry indicates that about 20 percent of total bagasse can be said to be excess after fulfilling internal energy demands [34]. This excess bagasse is one of the variables that are important to transportation and supply chain optimization through FDEA.

$$Surplus\ Bagasse = Total\ Bagasse\ Production \times 0.20 \quad (3)$$

**Monthly Surplus Bagasse Distribution:** The seasonal nature of bagasse production is that sugarcane crushing is done mostly between October and March/April with January being the peak period, leading to variation in the availability of surplus bagasse seasonally [35]. To represent this variation, the overall surplus bagasse was apportioned into monthly values depending on the intensity of crushing with higher proportions to the peak months and lower proportions to the start and the end of the season. The given distribution allows optimizing the supply chain by properly modelling time-dependent bagasse availability. Monthly surplus bagasse calculated by using (Eq. (4)).

$$Monthly\ Surplus\ Bagasse = Total\ Surplus\ Bagasse\ \times Monthly\ Distribution\ \% \quad (4)$$

**Table 3** The monthly surplus bagasse distribution used in this as follows.

| **Month** | **Surplus Bagasse Distribution** |
|---|---|
| October | 5% |
| November | 10% |
| December | 25% |
| January | 30% |
| February | 15% |
| March | 10% |
| April | 5% |

**Plant-to-Mill Distance Calculation:** Distance is a key variable which determines the cost of transportation, fuel consumption, efficiency of the entire supply chain. As there were no measurements of the actual distances between plants and mills, an estimation method was used based on district-to-district distances available in the literature [36] and intra-district distances based on the Haversine method. Furthermore, several production plants of bioethanol were presupposed in each district to create several pathways of supply chains to be assessed by FDEA. The overall distance between plants and mill was estimated by the following equation (Eq. (5) and (6)).

Total Distance = Intra-District Distance **(5)**
(If the plant and mill are in the same district)

Total Distance = District-to-District Distance + Intra-District Distance **(6)**
(If the plant and mill are in different districts)

This method gives a realistically close value of the transportation distance which is further employed in calculating costs and efficiency in the supply chain model.

**Diesel Price per Liter:** Diesel price is a key factor influencing transportation cost in the biomass supply chain. Historical diesel prices by district (20172025) were gathered using an official source tracking fuel prices [37] and allocated to each data record by the respective district and crushing season year. Where there were missing year values, data were averaged by district, whereas after the most recent year of available data, the most recent price (2025) was used. This variable was also used to calculate diesel cost per kilometre, total transportation cost, and transportation cost per ton of bagasse in optimizing the supply chain based on FDEA.

**Vehicle Type Selection:** The choice of vehicle is important in the bagasse supply chain in that it dictates transportation, consumption of fuel, and efficiency. According to industry practice and literature, short distance transportation is an appropriate use of tractors (50-75 HP with trolleys) whereas long distance transportation should use trucks (16-25 tonne capacity) because of the increased payload capacity [37]. Vehicle type was distributed in this study depending on transportation distance, to represent realistic biomass logistics situations.

If Distance $\leq 100$ km: Tractor
If Distance $> 100$ km: Truck

This categorization was also utilized in giving vehicle-specific details like mileage, payload capacity and cost of transportation to model the supply chain and conduct a FDEA efficiency analysis.

**Vehicle Payload Capacity:** Vehicle payload capacity is an important factor in transportation since it dictates the amount of bagasse that a vehicle can carry at a time, which in turn impacts the fuel consumption and costs per pound. The payload capacity used in this investigation was allocated depending on the type of vehicle and the distance of transportation as an indication of realistic logistics where a vehicle with a large capacity is favored over shorter distances. Trucks had a capacity of between 9 and 25 tonnes and the tractors had a capacity of between 5 and 15 tonnes. This distance-based capacity assignment enhances the efficiency of the supply chain by minimizing the number of trips and total transportation cost, and is also applied in the computation of trips, fuel consumption, and cost parameters based on efficiency assessment of FDEA.

**Vehicle Mileage (km per Liter):** Vehicle mileage is a vital parameter to estimate the use of diesel in the supply chain and transportation cost. In this analysis, the mileage was allocated on the basis of type of vehicle (truck or tractor) and size of payloads keeping in mind that the heavier the truck or tractor, the less fuel is used. Trucks that carried heavier loads (9 to 25 tonnes) received lower mileages whereas the tractors (5 to 15 tonnes) received mileages depending on the trailer setups and the load condition. This inverse correlation between payload and mileage is the actual practice of transportation in the real world, and was applied in the calculation of diesel consumption, transportation cost and cost per ton of bagasse to analyze efficiency of FDEA-based analysis of efficiency.

**Transportation Cost:** Calculation The parameters of computation of transportation cost were computed based on the diesel price, vehicle mileage, distance, and payload capacity to estimate the supply chain cost effectively. Diesel cost per kilometre is the cost of fuel per kilometre travelled whereas the total transportation cost is in relation to the distance between mill and plant. The cost per ton of bagasse transported was computed to determine the cost efficiency per unit of biomass transported.

$$Diesel\ Cost\ Per\ Km = \frac{Diesel\ Price\ Per\ Liter}{Vehicle\ Milage\ (kmpl)} \tag{7}$$

Transportation Cost=Total Distance × Diesel Cost per km (8)

$$Transportation\ cost\ per\ ton = \frac{Transportation\ cost}{Vehicle\ payload\ capacity} \tag{9}$$

### 3.4 Dataset Overview and Feature Description

The dataset used in the current research includes a complete package of characteristics, which combined together reflect the dynamics of sugar production, bagasse generation, logistics of transporting the same, and the efficiency of the entire supply chain. It has key identification characteristics like serial number, sub-serial number, mill code, and ownership type that enable orderly tracking and categorizing of records. The variables associated with production include the most important operational parameters such as the daily crush capacity, sugarcane crushed, sugar production, the percentage of sugar recovered and the dates of the crushing season (the beginning and the end of the sugar crushing season), which gives a detailed account of the mill-level productivity and performance. The bagasse related characteristics are very important in the measurement of the availability of biomass used in production of bioethanol. They are bagasse production percentage, bagasse production total, and surplus bagasse, which in totality have the generation and availability of lignocellulosic feedstock in operations of the supply chain. Moreover, the variables that are specific to plants (e.g., the name of a plant, plant district), as well as time-related variables (e.g., the month and the year), allow analyzing the flow of biomass within the supply chain both in space and season.

The modelling of supply chain efficiency revolves around transportation and cost-related parameters. Also among the key variables under this category are total transportation distance, diesel price per liter, vehicle type, vehicle payload capacity and vehicle mileage. These parameters are also employed to determine the critical cost parameters which include diesel cost per kilometre, total transportation cost and transportation cost per ton of bagasse. These obtained variables are the main inputs and outputs of the Fuzzy Data Envelopment Analysis (FDEA) model, which allows assessing the efficiency of routes at the level of routes. Moreover, the dataset includes monthly surplus bagasse and normalized FDEA efficiency score [38], the latter indicates the performance of any mill-to-plant transportation route in comparison to others. The fact that the operational and efficiency-based variables are included is a guarantee of a holistic representation of the biomass supply chain. Each of the features is well chosen with respect to the methodologies that had been applied in earlier peer-reviewed researches, thus making the data set robust and relevant to modelling, optimisation and prediction of biomass supply chain systems.

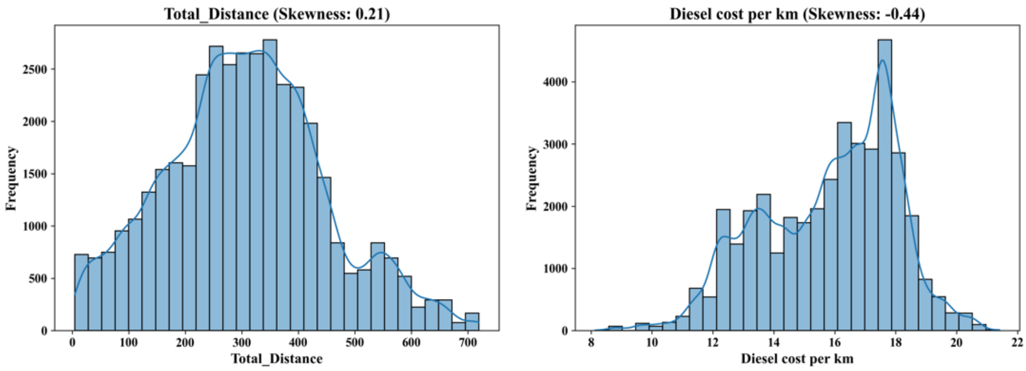
### 3.5 Exploratory Data Analysis

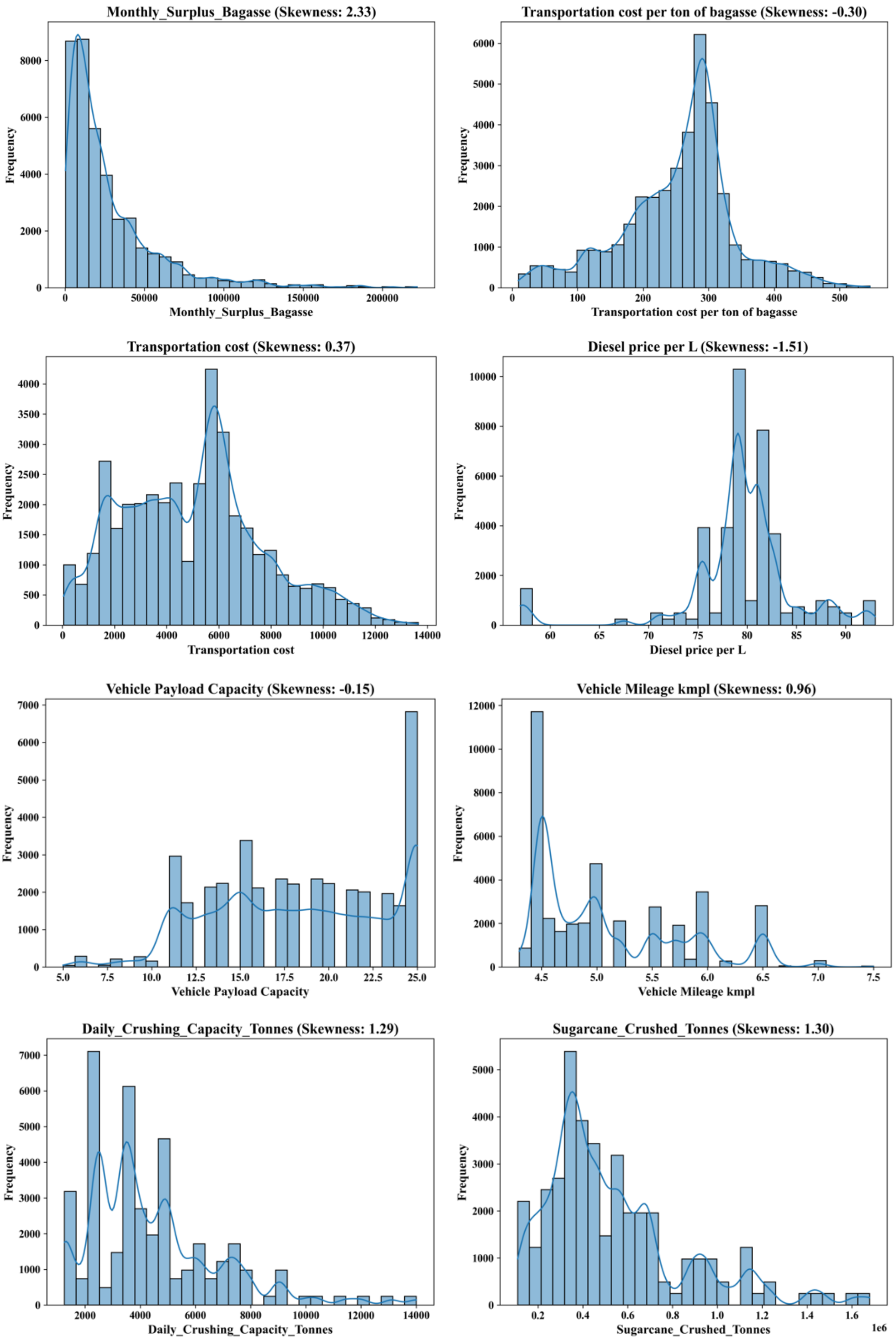
Exploratory Data Analysis (EDA) is an essential process in any data-driven modelling that entails the process of analysing and visualizing data with the aim of understanding the underlying data patterns, data distribution, and variable relationships prior to the application of optimization and machine learning algorithms. EDA assists in determining the skewness of the data, dependencies between features, and possible problems, which can greatly affect the performance of the model. In the given study, EDA was conducted to learn how the most important variables are distributed and how they impact the supply chain performance. Data distribution was analyzed through the histogram and skewness was detected, whereas the correlation analysis with the help of a heatmap was performed to recognize relationships between features, which are presented in Fig. 2 and Fig. 3 respectively.. They were useful in identifying the right variables and appropriateness of data to the FDEA and machine learning models, and hence enhanced the credibility of the supply chain optimization process [39][40], equation number 10 and 11 were used to derive Skewness and Pearson correlation coefficient respectively.

$$Skewness = \frac{E[\,(x - \mu)^3]}{\sigma^3} \tag{10}$$

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} \tag{11}$$

Fig. 2 shows the histogram distributions of the key input variables used in this study that gives us a complete picture of the statistical characteristics of these variables. The visualization aids in a detailed analysis of data distribution patterns, such as central tendency, dispersion, as well as skewness of variables like transportation distance, diesel cost, bagasse availability, production parameters, and logistics-related aspects. The skewness values observed, suggest the existence of both positive and negative skewed distributions which are inherent to the variability and non-uniformity of real-world biomass supply chain data. As an example, monthly surplus bagasse or production-related variables have right-skewed distributions, indicating that they contain high-value outliers, whereas cost-related variables are more or slightly asymmetric. This exploratory data analysis is crucial in justifying the quality of data, possible anomalies, and whether it is suitable to proceed with the further modelling processes, such as Fuzzy Data Envelopment Analysis (FDEA) and machine learning-based prediction. In addition, the knowledge of the distributional characteristics of the data aids in making informed choices on normalization, transformation and model choice to improve the quality of the proposed supply chain optimization framework in terms of robustness and reliability.




Monthly_Surplus_Bagasse (Skewness: 2.33)
Frequency
Monthly_Surplus_Bagasse
Transportation cost per ton of bagasse (Skewness: -0.30)
Frequency
Transportation cost per ton of bagasse
Transportation cost (Skewness: 0.37)
Frequency
Transportation cost
Diesel price per L (Skewness: -1.51)
Frequency
Diesel price per L
Vehicle Payload Capacity (Skewness: -0.15)
Frequency
Vehicle Payload Capacity
Vehicle Mileage kmpl (Skewness: 0.96)
Frequency
Vehicle Mileage kmpl
Daily_Crushing_Capacity_Tonnes (Skewness: 1.29)
Frequency
Daily_Crushing_Capacity_Tonnes
Sugarcane_Crushed_Tonnes (Skewness: 1.30)
Frequency
Sugarcane_Crushed_Tonnes
1e6

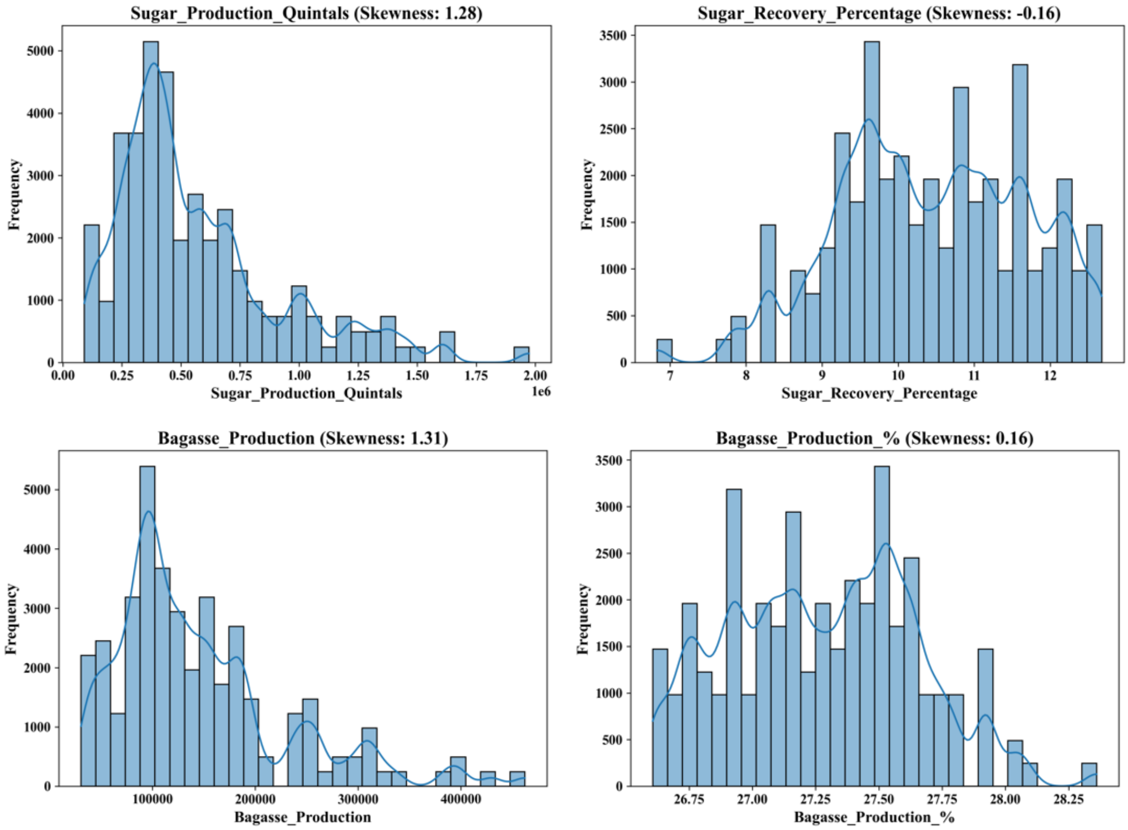



**Fig. 1.** Data distribution using histogram

Fig. 1 shows the heatmap of the correlation that presents the relationship between input variables and the desired FDEA efficiency score. The heatmap is a detailed visualization of the strength and direction of the linear relationships between features whose correlation coefficients have a range between -1 and +1 with positive coefficients representing direct relationships and negative coefficients representing inverse relationships. In the analysis, production-related variables, including sugarcane crushed, sugar production, and bagasse production are highly correlated, thus, showing that they are interdependent to each other in the process of biomass generation. On the same note, variables with respect to transportation such as total distance and transportation cost have high positive correlation, indicating that they show a direct proportional relationship. On the other hand, vehicle mileage is negatively correlated with cost related parameters, which reflects its antagonistic impact on transportation efficiency. Notably, the FDEA efficiency measure exhibits moderate associations with the key parameters of operations and logistics, which underscores the role of production capacity and transportation parameters in overall supply chain efficiency. This correlation analysis is important to feature selection and multicollinearity tests as well as model building in order to determine which variables are the most significant in predicting efficiency.

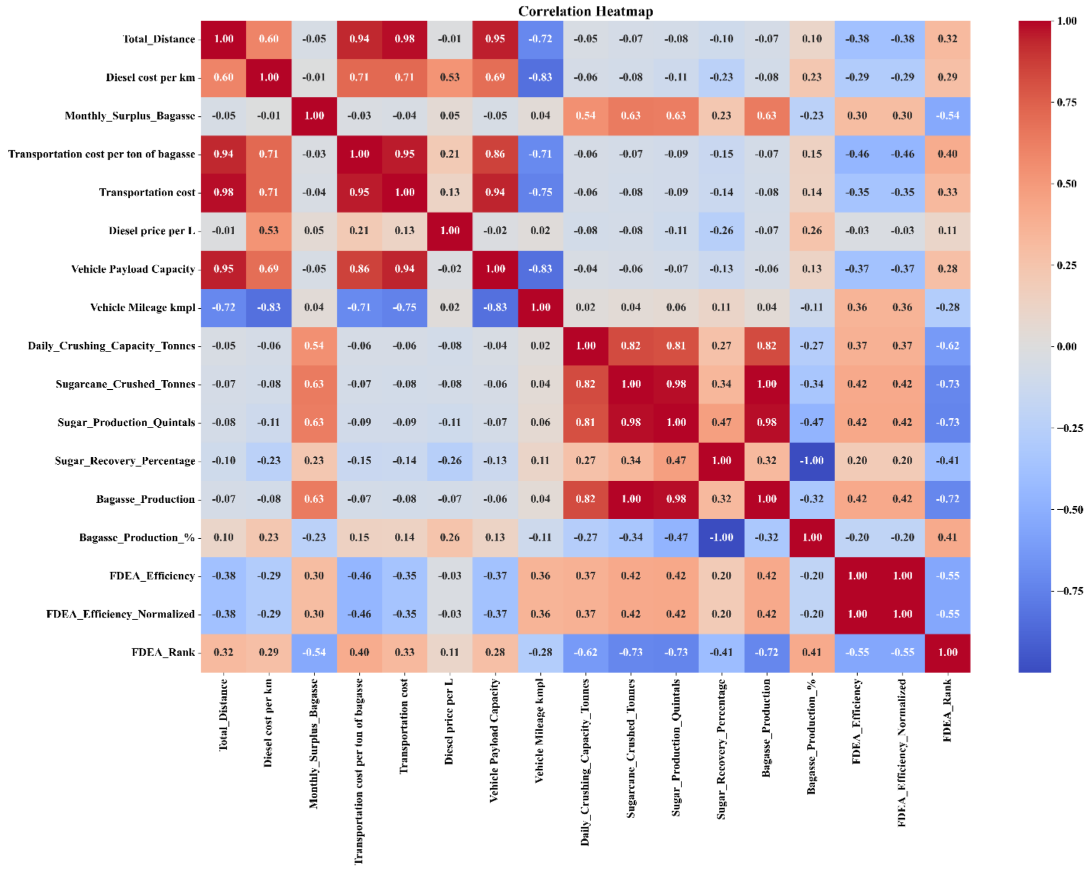



**Fig. 2** Correlation Heatmap

### 3.6 Fuzzy Data Envelopment Analysis (FDEA) Model

Fuzzy Data Envelopment Analysis (FDEA) is a development of the traditional DEA that is applied in the assessment of the relative efficiency of decision-making units (DMUs) in conditions of uncertainty with the inclusion of fuzzy inputs and outputs. It is extensively used in optimization of the supply chain to evaluate performance where various conflicting parameters like cost, distance, and resource usage are involved. Applied to biomass supply chains, FDEA can be used to identify efficient transportation paths, optimizing inputs (e.g. transportation cost, fuel consumption, distance) and outputs desired (e.g. bagasse supply) at the same time, uncertainties and variability in data. Because of its ability to simulate real-world inexactness and multi-criteria decision-making, FDEA can be an efficient instrument to analyze and optimize the biomass-to-bioethanol supply chain networks, as shown in recent research [41].

The Fuzzy Data Envelopment Analysis (FDEA) model was applied to assess and optimize the efficiency of bagasse transportation routes between sugar mills and bioethanol plants considering uncertainty. Under this method, a Decision-Making Unit (DMU) is a Plant-Mill-Month combination, and this allows the evaluation of the performance of a supply chain to be time-sensitive. Total distance and diesel cost per kilometre were chosen as the relevant input variables to be minimized, and monthly surplus bagasse as the desirable output, and transportation cost per ton of bagasse as the undesirable output to be minimized. All the inputs and outputs were reduced to a common scale of between 0 and 1 in order to make the DMUs comparable.

$$Normalized\ Value = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{12}$$

Natural Resources for Human Health
eISSN: 2583-1194

The range-based weights of desirable and undesirable outputs were then computed in order to represent their comparison of importance in the efficiency analysis (Eq. 1314). Using these weights, an objective function was developed so as to maximize desirable output and minimize undesirable output at the same time (Eq. 15). The DEA model has been developed as a linear programming problem that included input, desirable output, and undesirable output constraints to guarantee that DMUs can be feasibly and fairly compared. The efficiency score of any DMU was then calculated by maximizing the objective function with the specified constraints (Eq. 16).

Slack variables were also computed to further evaluate inefficiency, excess of input utilization and failure to achieve desired output relative to the efficient frontier (Eq. 17-20). Because the transportation parameters in the real world, including distance and cost of diesel, have uncertainties, triangular fuzzy numbers were used to establish fuzzy lower boundaries (Eq. 21-24). The efficiency calculation was then adjusted to these fuzzy inputs to come up with a more robust and realistic FDEA efficiency score (Eq. 25).

The efficiency scores were calculated and normalized between 0 to 1, and ranked to determine the most efficient mill-to-plant transportation routes. This combined FDEA framework allows making sound judgments in choosing routes that have the least transportation cost and less fuel usage and the most bagasse availability to enhance the overall efficiency and sustainability of the biomass supply chain.

$$R_g = \frac{1}{(m+s+h)(g_{max} - g_{min})} \tag{13}$$

$$R_b = \frac{1}{(m+s+h)(b_{max} - b_{min})} \tag{14}$$

$$Max \;\; \mathrm{Z} = R_g y - R_b b \tag{15}$$

$$Max \text{ Efficiency} = u y_i - w b_i \tag{16}$$

$$Slack_{Distance} = Actual\ Distance - Minimum\ Distance \tag{17}$$

$$Slack_{Diesel} = Actual\ Diesel\ Cost - Minimum\ Diesel\ Cost \tag{18}$$

$$\text{Slack Output} = \text{Maximum Bagasse - Actual Bagasse} \tag{19}$$

$$Total \text{ Slack} = Slack_{Distance} + Slack_{Diesel} + Slack_{Output} \tag{20}$$

$$Distance_L = 0.9 \times Distance_M \tag{21}$$

$$Distance_U = 1.1 \times Distance_M \tag{22}$$

$$Diesel_L = 0.95 \times Diesel_M \tag{23}$$

$$Diesel_U = 1.05 \times Diesel_M \tag{24}$$

$$FDEA \text{ Efficiency} = \frac{Efficiency_L + Efficiency_M + Efficiency_U}{3} \tag{25}$$

**3.7 Machine Learning Implementation**

In this section, machine learning models are applied to forecast the normalized FDEA efficiency score using principal operational, transportation and production-related variables of the bagasse supply chain. The input feature set consists of the total distance, diesel per kilometre, monthly surplus bagasse, cost of transportation, seat capacity of vehicles, vehicle mileage, cost of transportation per ton bagasse, sugarcane crushed, sugar production, bagasse production, and the time of year (month). The dependent variable is the normalized FDEA efficiency score. Multiplexed regression models were used to fit the data based on earlier research [42-45] and they include Linear Regression, Decision Tree Regressor, Support Vector Regressor (SVR), Polynomial Ridge Regression, AdaBoost Regressor and Random Forest Regressor. The choice of these models was to have a balanced representation of classical statistical methods, tree-based learning, kernel-based methods, and ensemble learning methods. These models can be broadly divided into non- ensemble and ensemble learning models. Non-ensemble models are directly learned on the data by a single predictive structure, but ensemble models combine several weak learners together to improve the predictive performance and robustness. All models were tested on standard regression performance measures such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), coefficient of determination ($R^2$), adjusted $R^2$, Mean Absolute Error (MAE) and explained variance score. These findings reveal that ensemble learning models tend to perform better than individual models, whereas more complex learning structures continue to do better in predictive accuracy by including complex relationships found in supply chain data.

*3.7.1 Non-Ensemble Models*

**a) Linear Regression**

A supervised learning method is called Linear Regression which models a relationship between a dependent variable and an independent variable or multiple independent variables. It approximates the most suitable linear equation by minimizing the error between the values observed and those predicted. The single input variable has the relationship as:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{26}$$

For multiple input variables, the model extends to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{27}$$

This model provides a baseline for evaluating linear dependencies between supply chain variables and efficiency scores.

**b) Decision Tree Regressor**

Decision Tree Regressor is a non-parametric supervised learning process, which divides the data into hierarchical subsets according to decision rules. The choice of each split is to reduce prediction error, which leads to homogeneous subsets. Standard Deviation Reduction (SDR), which is used to assess quality of a split, is:

$$SDR = SD(P) - \left(\frac{N_L}{N} SD(L) + \frac{N_R}{N} SD(R)\right) \tag{28}$$

An increased SDR implies a more optimal split, which results in better predictive performance. Decision trees are especially useful in the modelling of non-linear relationships in supply chain data.

**c) Support Vector Regressor (SVR)**

The Support Vector Regressor (SVR) is a generalization of the Support Vector Machines to regression problems; it attempts to estimate a functional within a specified tolerance (ε). It reduces the complexity of models and punishes the deviations outside the ε-insensitive region. The formulation of the optimization is presented as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$$

subject to: (29)

$$\begin{gathered} y_i - (w^T x_i + b) \leq \epsilon + \xi_i \\ (w^T x_i + b) - y_i \leq \epsilon + \xi_i^* \end{gathered}$$

SVR is robust to noise and outliers, making it suitable for real-world supply chain datasets.

**d) Polynomial Ridge Regression**

Polynomial Ridge Regression is a regression model based on a non-linear relationship with the use of L2 regularization and transforming features into polynomials to avoid overfitting. The objective function is defined to be:

$$\hat{\beta} = \arg\min_{\beta} \left( \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, \Phi_j(x_i))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{30}$$

The corresponding polynomial model is:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d \tag{31}$$

with coefficient estimation:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y \tag{32}$$

This approach balances model flexibility and generalization performance.

*3.7.2 Ensemble Models*

**a) AdaBoost Regressor**

An ensemble learning algorithm, AdaBoost (Adaptive Boosting) is an algorithm that increases the accuracy of prediction by combining many weak learners. The successive models are aimed at rectifying errors committed by the earlier models. The last prediction is provided by:

$$F(x) = \sum_{m=1}^{M} \alpha_m \, h_m(x) \tag{33}$$

The model minimizes exponential loss:

$$L = \sum_{i=1}^{N} \exp(-\,\alpha_m y_i h_m(x_i)) \tag{34}$$

This adaptive weighting mechanism enhances model performance in complex datasets.

**b) Random Forest Regressor**

Random Forest is an ensemble algorithm that builds numerous decision trees based on bootstrap sampling and random features. The resultant prediction is the average of the trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t\,(x) \tag{35}$$

Each tree predicts:

$$f_t(x) = \hat{c}_m \text{where } x \in R_m \tag{36}$$

Random Forest reduces overfitting and improves generalization by leveraging multiple independent models.

*3.7.3 Data Normalization*

The Standard Scaler method was used to ensure that features are scaled uniformly. This method normalizes the features, ignoring the mean and normalizing to unit variance:

$$x' = \frac{x - \mu}{\sigma} \tag{37}$$

where $\mu$ and $\sigma$ represent the mean and standard deviation, respectively. This normalization improves model convergence and stability.

*3.7.4 Performance Evaluation Metrics*

Model performance was evaluated using multiple regression metrics to ensure comprehensive assessment:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{38}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{39}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{40}$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n-1)}{n-p-1}\right) \tag{41}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{42}$$

$$EVS = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{43}$$

These metrics collectively evaluate prediction accuracy, error magnitude, and model explanatory power.

### 3.8 Custom Sequential Model Implementation

A custom sequential neural network model is a feedforward deep learning structure where several layers are stacked sequentially, with each layer being able to act and transform the output of the previous layer. This hierarchical form allows the model to get to know sophisticated and non-linear feature faces of supply chain data. The current research designed its own custom sequential neural network models to conduct a comparative analysis with the traditional machine learning algorithms to predict the normalized FDEA efficiency score. Two popular deep learning models, TensorFlow and PyTorch, were used to develop models to be robust and flexible to implement. The use of two frameworks allows to check the consistency and reliability of the models in various computational settings. Hyperparameter tuning was done in a systematic manner to tune the performance of the models such as variations in the key parameters like learning rate, training epochs, drop out rate, and learning rate scheduling strategies. These optimizations were critical to improve convergence behaviour, prevent overfitting, and the ability to generalize.

The models were trained and assessed based on a broad range of regression performance indicators such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R 2 ), adjusted R 2 and explained variance score. The adoption of various evaluation measures can guarantee a sound measure of predictive accuracy, error distribution, and model reliability. Table 6 provides an overview of the architectural design and hyperparameter settings of the TensorFlow and PyTorch-based sequential models in a comparative manner. The consistency of the model structure and training strategies used to make sure that the performance of the two implementations can be compared fairly and without bias is pointed out in the table.

**Table 4** Comparison of Hyperparameters of TensorFlow and PyTorch Sequential Models

| **Parameter** | **TensorFlow & PyTorch Sequential Models** |
|---|---|
| Framework | TensorFlow and PyTorch |
| Model Type | Sequential Neural Network (Feedforward Architecture) |
| Input Layer | TensorFlow: Dense (Input: 512)<br>PyTorch: Linear (Input: 512) |
| Hidden Layer 1 | TensorFlow: Dense (512) + Batch Normalization + ReLU<br>PyTorch: Linear (512) + BatchNorm1d + ReLU |

| Parameter | TensorFlow & PyTorch Sequential Models |
|---|---|
| Hidden Layer 2 | TensorFlow: Dense (256) + Batch Normalization + ReLU PyTorch: Linear (256) + BatchNorm1d + ReLU |
| Hidden Layer 3 | TensorFlow: Dense (128, ReLU) PyTorch: Linear (128) + ReLU |
| Dropout Rate | 0.2 (applied in both models) |
| Output Layer | TensorFlow: Dense (1) PyTorch: Linear (64 → 1) |
| Activation Function | ReLU |
| Loss Function | Mean Squared Error (MSE) |
| Optimizer | Adam |
| Learning Rate | 0.001 |
| Epochs | 60 |
| Validation Method | TensorFlow: Train–Validation Split PyTorch: Train–Test Split |
| Early Stopping Patience | 50 |
| Best Model Restoration | Enabled in both (restore best weights/model state) |
| Feature Scaling | Standard Scaler (applied) |
| Evaluation Metrics | TensorFlow: MSE, RMSE, MAE, $R^2$, Adjusted $R^2$, Explained Variance PyTorch: MSE, $R^2$ |

This two-fold structure implementation achieves a high-quality methodology and allows the deep-learning performance to be completely compared with the traditional machine-learning frameworks, which consequently enhances the predictive power of the presented supply chain optimization framework.

## 4. RESULTS AND DISCUSSION

The normalized FDEA efficiency score was systematically estimated using the predictive performance of multiple machine learning and deep learning models to estimate the predictive performance of these models. The comparative analysis reveals the usefulness of both traditional regression models and modern neural network architectures in the ability to reflect the underlying relationships in the biomass supply chain data set. Ensemble and tree-based models were found to perform better than the linear models, which are among the machine learning models. AdaBoost Regressor model was found to be the most accurate of traditional models with an $R^2$=0.950, closely followed by the Decision Tree Regressor, and the Random Forest Regressor ($R^2$=0.947, and $R^2$=0.943). Polynomial Ridge Regression was also found to be an effective model with high predictive ability with an $R^2$ value of 0.930, which shows that it is effective in non-linear relationship modelling using the expansion and regularization of features in the form of a poly-nomial. Although these models performed well, the custom deep learning architectures implemented with PyTorch and TensorFlow were much better than all traditional machine learning methods. The PyTorch model attained a high $R^2$ of 0.975 with a further increase of the prediction accuracy in the TensorFlow model with an $R^2$ score of 0.978. It can be improved by the fact that deep neural networks can learn multifaceted, high-dimensional, non-linear interdependences among variables in a supply chain, such as transportation cost, production capacity and biomass availability. The outperformance of deep learning models suggests that the dependence between input features and FDEA efficiency are non-linear and multi-dimensional and can not be sufficiently represented by conventional regression methods. Moreover, incorporating various hidden layers, batch normalization and dropout regularization into the tailor made sequential models improves the overall performance of generalization and minimizes the chances of overfitting.

**4.1 FDEA**

*4.1.1 FDEA-Based Supply Chain Efficiency Analysis (Case Study: January 2025)*

This part provides a case study-based Fuzzy Data Envelopment Analysis (FDEA) of the biomass supply chain network of the Western Maharashtra region covering six major districts including Pune, Kolhapur, Satara, Sangli, Solapur and Ahmednagar in the month of January 2025. The main aim of the analysis is to assess the efficiency of mill-to-plant transportation routes and determine the best routes to reduce the cost of logistics and as much as possible use the excess sugarcane bagasse to produce bioethanol. The individual sugar mills and bioethanol plants are modeled as a Decision Making Unit (DMU), each representing a transportation route. In this study, 17 DMUs were taken into account and each one of them was characterized by transportation distance, fuel cost, production parameters, and bagasse availability. This case study is a realistic example of regional biomass supply chain operations, which is the starting point of further predictive modelling based on machine learning methods.

A multi-input and multi-output model was used to develop the FDEA model to identify the trade-offs between minimization of transportation cost and maximization of biomass utilisation [46-48]. Parameters in transportation such as total distance, cost of diesel per kilometre and transportation cost per ton of bagasse were regarded as input variables to be minimized since they are the direct causes of inefficiency in logistics. On the other hand, the output variables were included like monthly excess bagasse, allocated bagasse to reflect the favorable results of the system, which symbolizes good use of biomass and productivity of the supply chain. This systematic development allows a thorough efficiency analysis by taking into consideration the cost of operation and the flow of biomass at the same time and hence offering a well-balanced evaluation of the performances of the supply chains of the various DMUs.

*4.1.2 FDEA Efficiency Results*

The outcomes of the FDEA analysis show that there is a high level of variation in efficiency of the investigated DMUs, and this implies that there is heterogeneity in the supply chain performance in the Western Maharashtra region. The normalized efficiency scores are found to be between 0.072 to 0.465 and this indicates that there are efficient and inefficient transportation routes. The route that recorded the most efficiency was the Satara_Plant_3_15201_January with a score of 0.465, Solapur_Plant_5_69016_January and Satara_Plant_3_38801_January with efficiency of 0.407 and 0.390 respectively. These are the paths with shorter transport distances and comparatively low cost per ton, which allows the efficient transportation of biomass. Conversely, Ahmednagar_Plant_4_63301_jan and Kolhapur_Plant_1_17201_jan had the lowest efficiency values of 0.072 and 0.081, respectively. These paths imply much longer transportation routes, more than 140 km, and high logistics expenses, which leads to low overall productivity despite a high supply of bagasse. The identified difference proves the paramount importance of transportation parameters in the supply chain efficiency.

**Table 5** Comparative Analysis of DMU's for the Western Maharashtra Region

| DMU_ID | Total_Distance in Km | Diesel cost per km in Rs. | Monthly_Surplus_Bagasse in Tons | Transportation cost per ton of bagasse in Rs. | FDEA_Efficiency_Normalized | FDEA_Rank | Allocated_Bagasse |
|---|---|---|---|---|---|---|---|
| Pune_Plant_4_60701 | 36.73 | 13.16625 | 188954.52 | 48.36125 | 0.26168 | 22 | 188955 |
| Pune_Plant_4_13601 | 49.95 | 13.62125 | 143502 | 61.8525 | 0.15789 | 49 | 121522 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Kolhapur_Plant_1_50801 | 13.58 | 11.30375 | 58794.72 | 25.58375 | 0.16679 | 46 | 58794.7 |
| Kolhapur_Plant_1_18001 | 13.58 | 11.06142 | 43752.48 | 25.03571429 | 0.12765 | 75.5 | 43752.5 |
| Kolhapur_Plant_1_17201 | 164.48 | 13.18875 | 206241.84 | 180.7725 | 0.08133 | 165 | 45329.8 |
| Satara_Plant_3_15201 | 10.08 | 11.57125 | 145767.78 | 19.44 | 0.46532 | 2.5 | 145768 |
| Satara_Plant_3_38801 | 20.02 | 12.46125 | 184344.24 | 31.18375 | 0.39053 | 6 | 184344 |
| Satara_Plant_3_69023 | 20.02 | 12.46125 | 92877.12 | 31.18375 | 0.20084 | 32 | 20325 |
| Sangli_Plant_2_15401 | 30.09 | 12.74375 | 111544.62 | 42.6075 | 0.18169 | 40 | 111545 |
| Sangli_Plant_2_46901 | 7.63 | 11.32714 | 38355.78 | 14.40285714 | 0.15305 | 57 | 9310.38 |
| Solapur_Plant_5_69016 | 4.46 | 10.99875 | 88525.2 | 9.81125 | 0.40721 | 5 | 88525.2 |
| Solapur_Plant_5_52001 | 32.69 | 13.06375 | 221821.08 | 47.45 | 0.33156 | 10 | 221821 |
| Solapur_Plant_5_17801 | 32.69 | 13.06375 | 125021.52 | 47.45 | 0.18897 | 35 | 80220.7 |
| Ahmednagar_Plant_4_12701 | 27.47 | 12.205 | 87740.64 | 37.25166667 | 0.15703 | 51 | 87740.6 |
| Ahmednagar_Plant_4_66001 | 18.48 | 11.64166667 | 59127.72 | 30.735 | 0.13913 | 70 | 59127.7 |
| Ahmednagar_Plant_4_13101 | 56.03 | 13.672 | 93859.56 | 63.836 | 0.09787 | 119.5 | 93859.6 |
| Ahmednagar_Plant_4_63301 | 146.28 | 12.155 | 162219.6 | 161.64 | 0.07211 | 212 | 69839.1 |

The comparative study of the DMUs can shed more light on the factors that determine the efficiency performance. The most efficient DMUs are mostly found in the Satara and Solapur districts with distances of transportation of 4.46 km to 20.02 km and a cost per ton ranging between 9.81 to 31.18. All these positive conditions can lead to increased efficiency scores, which are over 0.33 in the majority of cases. Conversely, the least efficient DMUs, most of which belong to Ahmednagar and Kolhapur districts are linked with greater transportation distances, up to 164.48 km, and much greater cost per ton values, over 160. This sharp contrast shows that the main factors that determine the efficiency of supply chain are transportation distance and logistics cost. Importantly, even DMUs that have high surplus bagasse cannot be more efficient in the case of unfavorable transportation conditions, and the strategy of selecting routes and plant location should be optimized.

*4.1.4 District-Level Insights*

The analysis at the district level indicates clear spatial trends of the efficiency of supply chains in the study area. The efficiency level of Satara and Solapur districts is always high with geographical proximity of both sugar mills and bioethanol plants in favorable geographical location leading to a shorter transportation distance and less fuel consumption. Conversely, Ahmednagar and Kolhapur districts have lower efficiency indices, mainly because of the lengthy transport distances and higher logistic expenses. This geographical gap highlights the importance of regional planning and optimization of infrastructure to enhance the performance of biomass supply chains. The results indicate that decentralization of processing facilities or moving bioethanol facilities away could be an effective way to improve the performance of the low-performing districts.

*4.1.5 Effect of Key Variables on Efficiency.*

The analysis also shows that the most significant variables that impact FDEA efficiency are transportation distance and cost per ton. DMUs that have a lower transportation distance are always more efficient with higher scores whereas those that have longer transportation paths have a drastic decrease in their performance because of higher consumption of fuel and higher cost of operation. Even though excess bagasse has a positive effect on the output performance, its effect is less significant to that of transportation factors. This implies that an effective logistics planning is more critical than the volume of production in optimizing supply chain performance. The findings also confirm the correlation analysis that was conducted above with strong relationships found between the cost, distance and efficiency variables.

The distribution of the allocation outcomes of the FDEA model shows that there is a preference towards the efficient transportation lines in biomass distribution. DMUs which are highly efficient are allocated as much or as near as possible of surplus bagasse, as in the case of Satara_Plant_3_15201_jan and Solapur_Plant_5_52001_jan the allocation is quite close to the amount of surplus. Conversely, ineffective routes are allocated less meaning that the model penalizes well high-cost and long-distance routes. This allocation scheme will guarantee that biomass resources are used optimally and transportation costs are minimal hence improving on the supply chain performance. The findings affirm that FDEA is a potentially useful decision-support tool in bioethanol supply chains to assist in the allocation of resources and optimization of logistics.

The comparative analysis that is provided in Table X shows how FDEA models have developed and been applied in the areas of supply chain and efficiency evaluation. The main attention of previous studies has been on the inclusion of uncertainty in efficiency measurement, the greater robustness with the help of fuzzy logic, and better decision-making with the assistance of enhanced DEA-related frameworks. An example is that Gilani et al. [13] and Tavassoli et al. [21] highlighted the applicability of FDEA in the sustainability and multi-stage supply chain optimization, and Masoud Vaseei et al. [22] and Hemmati et al. [20] showed the enhanced accuracy and discrimination capability using uncertainty-sensitive efficiency modelling.

**Table 6:** Comparison of FDEA-Based Models with Existing Literature

| Study | Key Findings |
|---|---|
| Gilani et al. [13] | Developed an FDEA-based supply chain model that enhanced decision-making under uncertainty by enabling optimal selection of supply points and sustainable network design. The model outperformed deterministic approaches in terms of robustness, particularly in balancing economic, environmental, and social objectives. |
| Masoud Vaseei et al. [22] | Integrated DEA with bootstrap simulation to improve the reliability of efficiency evaluation under uncertainty. The results revealed significant variation in efficient DMUs, demonstrating that uncertainty-aware models provide more realistic efficiency assessment and improved identification of inefficient units. |
| Hemmati et al. [20] | Proposed an FDEA model for supplier evaluation that generated interval-based efficiency scores and robust rankings. The study highlighted improved discrimination power and reliability compared to traditional DEA approaches. |

| Study | Key Findings |
|---|---|
| Tavassoli et al. [21] | Introduced a fuzzy network DEA (FNDEA) model for multi-stage supply chains, enabling fair efficiency decomposition under uncertainty. The model ensured balanced efficiency evaluation across stages by incorporating optimistic and pessimistic efficiency bounds. |
| **Present Study** | Developed an FDEA-based framework for evaluating bagasse supply chain efficiency from sugar mills to bioethanol plants. Efficiency scores ranged from 0 to 1, where higher values indicate routes with lower transportation distance and diesel cost and higher biomass utilization. The model effectively identified cost-efficient and optimized transportation routes, providing a practical decision-support tool for real-world supply chain optimization. |

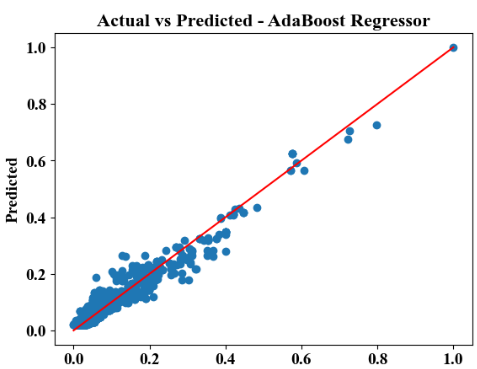
Nevertheless, the majority of the current research is either theoretical or generalized on the application of supply chain without considering the real-world biomass logistics systems. The current paper, in contrast, expands the use of FDEA to a practical and local context of bagasse transportation within the bioethanol supply chain of Western Maharashtra. In contrast to earlier methods, this research incorporates the transportation distance, the cost of fuel, and the distribution of the biomass into one coherent efficiency model, allowing to optimize routes at the route level under the conditions of actual operations. Moreover, the findings of this paper are used as a basis to future machine learning-based prediction, thus closing the gap between predictive analytics and optimization.

Although FDEA offers a powerful framework to assess the efficiency of supply-chain, it demands that optimization models are solved on a case-by-case basis of DMU, thus making it potentially challenging to scale to large-scale applications. To address this shortcoming, machine learning models are created in the following section to forecast FDEA efficiency scores using input variables. This integration allows the efficiency to be estimated rapidly and at scale without requiring the optimization to be repeated, and thus allows real-time decisions and real-time optimization of the supply chain to be made. The shift between FDEA and machine learning creates a hybrid framework that integrates efficiency assessment with predictive modelling, which predisposes the greater applicability of the proposed system.

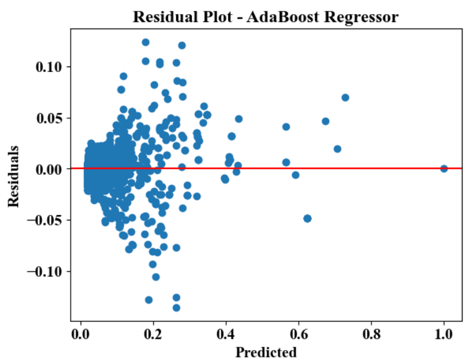
### 4.2 Comparative Analysis of Machine Learning and Deep Learning Models

#### *4.2.1 Machine Learning*

AdaBoost Regressor showed good predictive performance with a minimum MSE of 0.00011, RMSE of 0.011, $R^2$ of 0.953, adjusted $R^2$ of 0.953, MAE of 0.006 and a value of the explained variance of 0.953. The actual vs. predicted plot (Fig. 4a) shows that there is a high level of alignment along the 45° reference line and this confirms the correct prediction through the dataset. The residual plot (Fig. 4b) indicates that there is no systematic bias because the error values are randomly distributed around the centre with little dispersion and a high level of generalization of the model.



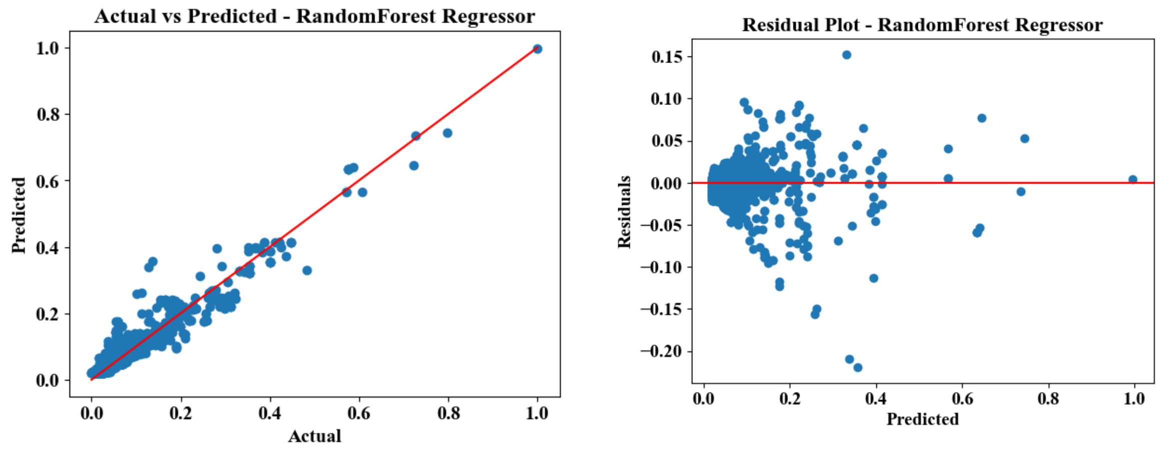

a. Actual vs Predicted plot




b. Residual Plot

**Fig. 3.** Actual vs. predicted and residual plots for AdaBoost Regressor.

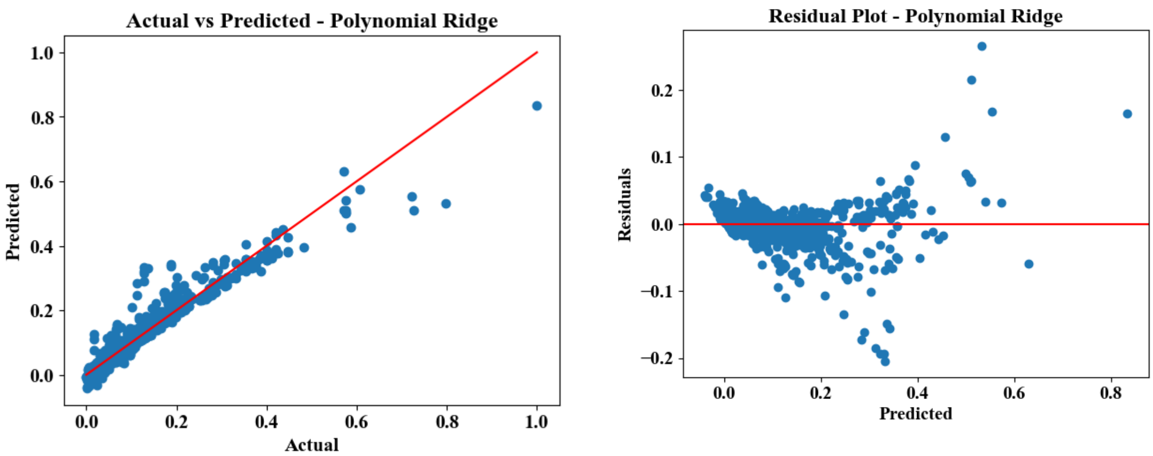
The Random Forest Regressor also demonstrated a high degree of predictive ability with an MSE of 0.00014, RMSE of 0.012, $R^2$ of 0.943, adjusted $R^2$ of 0.943, MAE of 0.006 and explained variance of 0.943. Predicted values closely follow actual values as indicated in Fig. 5a especially in dense data areas. The residual plot (Fig. 5b) shows that the distribution of the errors is rather homogeneous, but there is a slight increase in the variance on the range of mid-range predictions. The model is robust as a result of the ensemble averaging mechanism.




a. Actual vs Predicted plot b. Residual Plot

Fig. 4. Actual vs. predicted and residual plots for Random Forest Regressor.

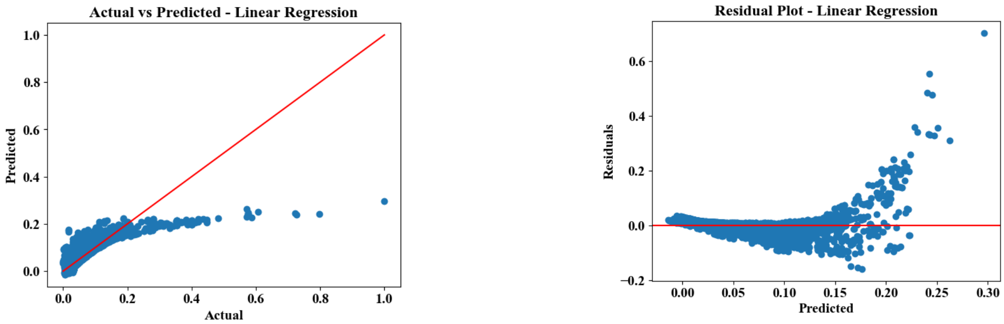
The Polynomial Ridge model attained a high predictive accuracy of MSE = 0.00017, RMSE = 0.013, MAE = 0.006 and good goodness-of-fit statistics ($R^2$ = 0.930, adjusted $R^2$ = 0.930, explained variance = 0.930). The actual vs. predicted plot (Fig. 6a) indicates that there is good agreement between the actual and predicted values of low and mid-range values; but some slight deviations appear at the high values of efficiency. The residual plot (Fig. 6b) shows that the dispersion is moderate, which means that the model is able to capture non-linear patterns, but it is less effective in accommodating extreme variations than the ensemble techniques.




a. Actual vs Predicted plot b. Residual Plot

**Fig. 5.** Actual vs. predicted and residual plots for Polynomial Ridge model

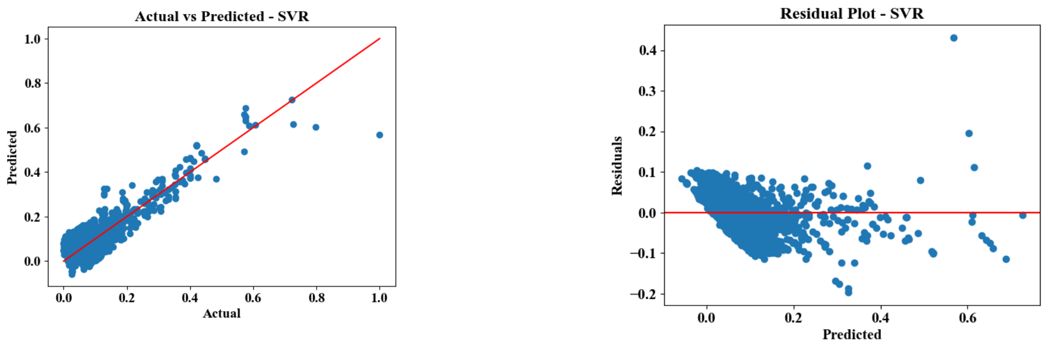
Linear Regression model had a moderate predictive performance with MSE=0.00091, RMSE=0.030, MAE=0.015 and $R^2$=0.620. The comparatively small $R^2$ value indicates that the model is not able to explain non-linear relationships in the data. As it can be seen in Fig. 7a, forecasting is far off the reality especially at the efficiency scores of above 0.3. The residual plot (Fig. 7b) indicates heteroscedasticity and model underfitting as it is increasing in value as predicted values are increasing.




a. Actual vs Predicted plot b. Residual Plot

**Fig. 6.** Actual vs. predicted and residual plots for Linear Regression model

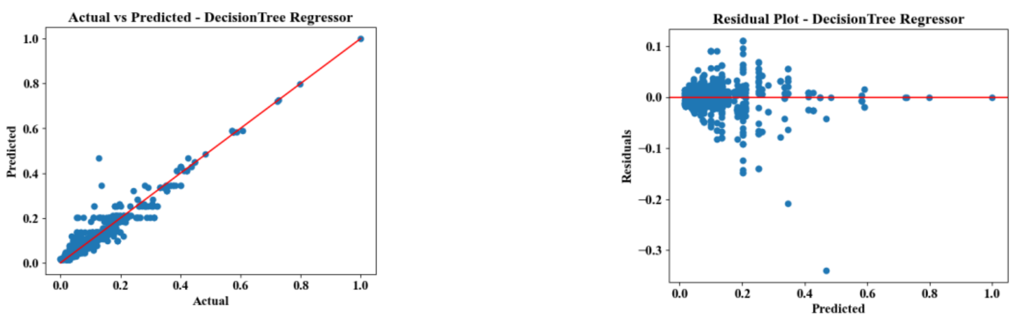
The comparatively low predictive power of the SVR model gave an MSE of 0.00331, RMSE of 0.058 and MAE of 0.363 with a value of $R^2$ of 0.417. The lower performance is mainly due to its poor ability to intercept more non-linear patterns at higher efficiency levels. Actual vs. predicted plot (Fig. 8a) indicates that there are huge deviations compared to the ideal prediction line particularly at values above 0.5. The residual plot (Fig. 8b) shows some observable dispersion and outliers, which renders the model unstable and with less generalization abilities.




a. Actual vs Predicted plot b. Residual Plot

**Fig. 7.** Actual vs. predicted and residual plots for SVR model

Decision Tree Regressor showed good results and the MSE, RMSE, MAE and $R^2$ are 0.00013, 0.011, 0.006, and 0.948 respectively. The predicted vs. actual plot (Fig. 9a) depicts that there existed a good agreement between the observed and predicted values, which proves the model to be good at identifying non-linear relationships. The residual plot (Fig. 9b) shows that, the majority of the residuals are concentrated around the value of zero and have little dispersion though there are few outliers because the model is sensitive to the partitioning of the data.




a. Actual vs Predicted plot b. Residual Plot

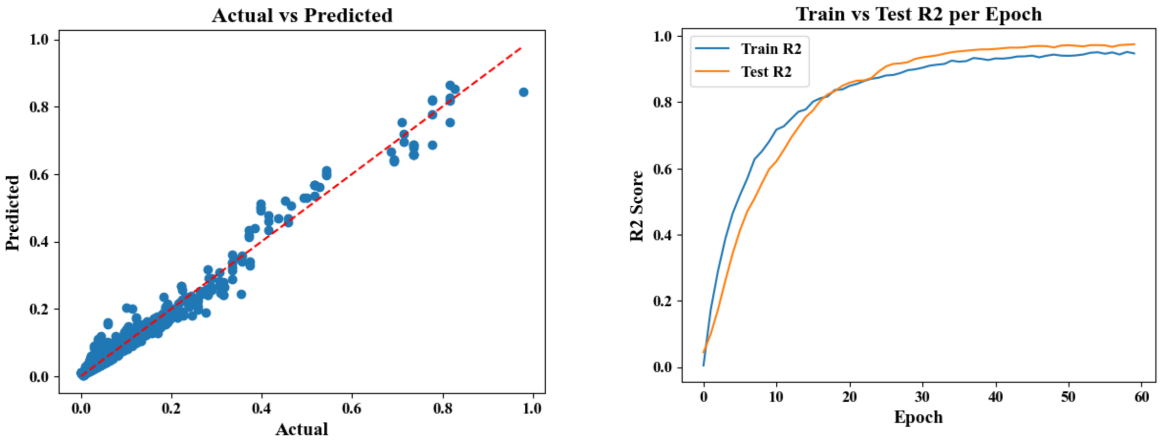
**Fig. 8.** Actual vs. predicted and residual plots for Decision Tree Regressor

Natural Resources for Human Health
eISSN: 2583-1194

The comparative analysis demonstrates that ensemble models (especially AdaBoost and Random Forest) are superior to conventional regression methods because they are able to recognize complex and non-linear relationships in the supply chain data set. Polynomial Ridge Regression proves to be a good performer as well, which implies that the data also has non-linear trends. Linear Regression and SVR, on the other hand, have a relatively lower accuracy, which indicates the weaknesses of simple linear assumptions and the use of kernels to model high-dimensional supply chain systems. The findings show clearly that the biomass supply chain efficiency issue is non-linear in nature and necessitates the use of sophisticated modelling methods to predict it.

*4.2.2 Deep Learning*

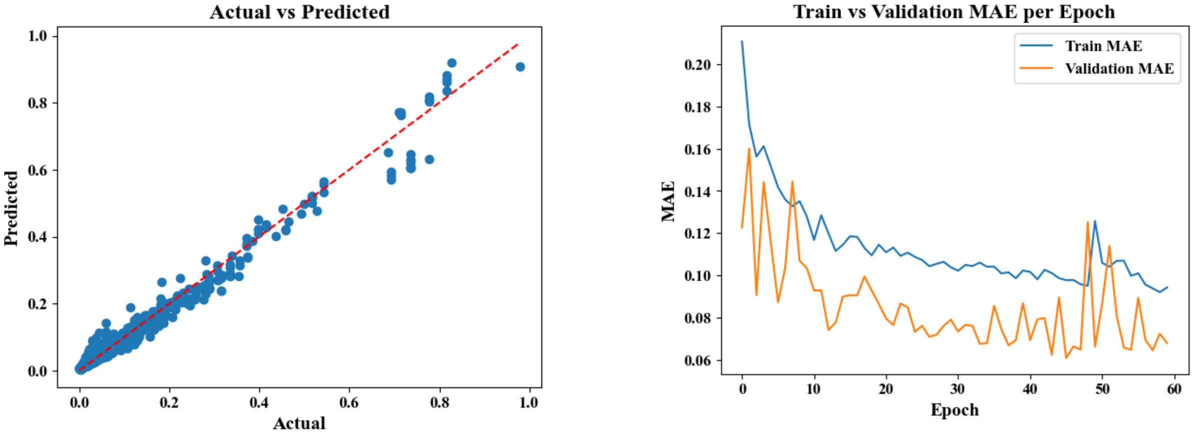
The custom sequential neural network realized in the PyTorch framework showed a better predictive performance with a Mean Squared Error (MSE) of 8.30 × $10^{-5}$, Root Mean Squared Error (RMSE) of 0.009, Mean Absolute Error (MAE) of 0.01 and coefficient of determination ($R^2$) of 0.975. The large $R^2$ value suggests that the model is powerful and capable of capturing the complex and non-linear interactions among the variables of the supply chain, and thus the model produces the most accurate predictions of the normalized FDEA efficiency score. The actual and predicted plot (Fig. 10a) shows the points stretched closely around the desired 45 degree reference line, which the model is very precise in predicting throughout the entire range of efficiency values. Moreover, the train vs. test $R^2$ convergence plot (Fig. 10b) illustrates that there is stable learning behaviour as both the training and testing curves converge without much expansion. This shows that the model can perform well in generalization and has little overfitting.




a. Actual vs Predicted plot b. Train vs. Test $R^2$ score plot

**Fig. 9.** (a) Actual vs. predicted values and (b) Train vs. test $R^2$ score per epoch for the PyTorch-based neural network model.

The overall most accurate of all the models considered the custom neural network model created with the TensorFlow framework, with an MSE of 7.10 × $10^{-5}$, RMSE of 0.008, MAE of 0.01 and a $R^2$ value of 0.978. These results highlight the effectiveness of the optimized deep learning architecture in modelling highly non-linear relationships present in the biomass supply chain dataset. The actual vs. predicted plot (Fig. 11a) depicts an even better correlation of the predictions with observed values than other models, which is better prediction performance in both low and high efficiency regimes. Also, the train vs. validation MAE plot (Fig. 11b) shows that there is a consistent decrease in error with epochs with little fluctuation, indicating that convergence is stable and that the model can learn generalized well. The lack of divergence between the training and validation curves is an additional reason to suggest that overfitting is well managed by regularization methods like dropout and early stopping.




a. Actual vs Predicted plot b. Train vs. Test R² score plot

**Fig. 10.** (a) Actual vs. predicted values and (b) Train vs. validation MAE per epoch for the TensorFlow-based neural network model

All machine learning (ML) and deep learning (DL) models are thoroughly compared in Tables 6, and their performance is observed in terms of conventional regression measures.

**Table 7:** Estimation of Evaluation Metrics for Machine Learning Models

| Model | MSE | RMSE | $R^2$ Score | Adjusted $R^2$ | MAE | Explained Variance |
|---|---|---|---|---|---|---|
| AdaBoost Regressor | 0.00011 | 0.011 | 0.953 | 0.953 | 0.006 | 0.953 |
| Random Forest Regressor | 0.00014 | 0.012 | 0.943 | 0.943 | 0.006 | 0.943 |
| Polynomial Ridge | 0.00017 | 0.013 | 0.930 | 0.930 | 0.006 | 0.930 |
| Linear Regression | 0.00091 | 0.030 | 0.621 | 0.620 | 0.015 | 0.621 |
| SVR | 0.00331 | 0.058 | 0.417 | 0.420 | 0.052 | 0.363 |
| Decision Tree Regressor | 0.00013 | 0.011 | 0.948 | 0.947 | 0.006 | 0.948 |
| Optimized PyTorch ANN | 8.30E-05 | 0.009 | 0.975 | 0.975 | 0.01 | 0.975 |
| Optimized TensorFlow ANN | 7.10E-05 | 0.008 | 0.978 | 0.978 | 0.01 | 0.979 |

The results obtained were compared to the existing literature on the topic to test the efficiency of the given framework. Table 7 includes a comparative study of various machine learning and deep learning methods that have been used in solving biomass and issues related to supply chains. The prior research work has mostly concentrated on either predictive modelling or optimization separately. As an example, Razmi et al., (2025) and Peng et al., (2023) to predict biomass and optimize logistics used machine learning models with R2s of 0.88-0.91. Equally, Hai et al., (2023) showed that the deep learning models are effective in bioenergy systems with performance metrics of approximately 0.93. Nevertheless, these papers are not connected with efficiency analysis models like Data Envelopment Analysis (DEA), which reduces their use in optimization of routes at the route level.

Other articles, like Han and Zhang (2023), used machine learning together with optimization methods but the uncertainty modelling or efficiency benchmarking. Similarly, Zhao et al., (2024) emphasized biomass prediction with the help of ML models, but did not deal with supply chain optimization or performance evaluation. Conversely, the current research suggests a new hybrid model that combines machine learning, fuzzy DEA and the deep learning methods in the optimization of biomass supply chain. The designed neural network, based on the TensorFlow, demonstrated the best predictive performance with an R2 of 0.978, compared to the existing methods. This has been enhanced by the integration of predictive modelling, efficiency analysis and uncertainty management in a single model. Moreover, in comparison with other investigations, the proposed solution can be used to assess the efficiency of routing at a route level, optimize costs, and make predictions in the presence of uncertainty, which makes it a more realistic application to the biomass supply chain in the real world. Thus, the findings indicate that the proposed ML-FDEA-based framework has a substantial improvement in the accuracy, robustness and practical applicability over the current methodologies.

**5. CONCLUSION**

The current research was able to come up with a Bagasse Supply Chain Management (SCM) model that has a purpose of maximizing transportation of biomass by ensuring that the most effective routes are taken. An extensive and organized dataset was built based on the main production and transportation and cost-related parameters, obtained and confirmed with the reports of FAO and previous peer-reviewed works. The suggested Fuzzy Data Envelopment Analysis (FDEA) model was successfully used to compare plant-mill transportation routes by having normalized efficiency scores of 0 to 1, with higher scores implying that a route delivers more bagasse with less transportation distance and diesel cost. These findings indicate that the model is effective as a sound decision-support model in the bioethanol supply chain planning process as it is reliable in identifying cost-effective and resource-optimal routes under the real-world conditions.

In order to further increase the analysis power of the framework, machine learning models were introduced to forecast FDEA efficiency scores. Ensemble learning models such as AdaBoost, Random Forest and Decision Tree had high predictive performance with R 2 of 0.950, 0.943 and 0.947 respectively and Polynomial Ridge Regression had an R 2 of 0.930. Conversely, non-ensemble models like Linear Regression (R2 = 0.62) and Support Vector Regression (R2 = 0.41) performed relatively worse, which underscores the inefficacy of less complex modelling methods in modeling the complex relationship between supply chains. In addition, tailor-made deep learning models written in PyTorch and TensorFlow had higher predictive precision with R 2 of 0.975 and 0.978 correspondingly, reflecting an improved predictive ability to capture non-linear and high-dimensional interactions in the data.

The comparison with the current literature attests to the fact that the machine learning models adopted are either better or on par with the reported results in the past, and specifically, the ensemble models. Equally, the suggested FDEA model proves to be consistent with the proven FDEA-based models in supply chain management, supplier assessment, and sustainable network design, thus justifying its strength in managing uncertainty, assessing efficiency, and prioritizing decision-making units in various fields of applications.

FDEA integrated with machine learning creates a hybrid framework, balancing the efficiency assessment with predictive analytics, providing a scalable and viable solution to biomass supply chain optimization. Future research can be aimed at adding dynamic parameters in real-time like traffic dynamics, season changes, and fuel prices changes to increase the accuracy of the model. Also, incorporation with GIS-driven routing systems and the creation of new hybrid models integrating deep learning with optimization methods can greatly enhance the decision-making process. The application of the proposed framework to other biomass resources and broad-scale industrial applications will make it even more scalable and applicable in real life.

**Nomenclature:**

| **Parameter** | **Description** |
|---|---|
| y | Predicted dependent variable |
| x | Independent feature |
| β0 | Intercept (value of y when x = 0) |
| β1 | Slope coefficient (change in y for one-unit change in x) |
| ε | Error term (difference between actual and predicted values) |
| y | Predicted dependent variable |

| | |
|---|---|
| x1, x2, …, xn | Independent features |
| β0 | Intercept |
| β1, β2, …, βn | Coefficients showing influence of each feature on y |
| ε | Error term |
| SDR | Standard Deviation Reduction (improvement after split) |
| SD(P) | Standard deviation of parent node |
| SD(L) | Standard deviation of left child node |
| SD(R) | Standard deviation of right child node |
| N | Total samples in parent node |
| N_L | Samples in left child node |
| N_R | Samples in right child node |
| C | Regularization parameter controlling trade-off |
| ξ_i, ξ_i* | Slack variables for ε-tube violations |
| N | Number of samples |
| F(x) | Final AdaBoost regression prediction |
| h_m(x) | Prediction from m-th weak learner |
| α_m | Weight assigned to weak learner |
| M | Total number of weak learners |
| L | Loss function |
| y_i | True value of sample i |
| h_m(x_i) | Predicted value of sample i |
| N | Number of data samples |
| ŷ | Final predicted output |
| T | Total number of trees |
| f_t(x) | Prediction from t-th decision tree |
| R_m | Region (leaf node) where input x belongs |
| ĉ_m | Mean target value in region R_m |
| $X$ | Original value |
| $X_{min}$ | Minimum value of that variable in the month |
| $X_{max}$ | Maximum value of that variable in the month |
| $R_g$ | Weight of desirable output |
| $m$ | Number of inputs |

| | |
|---|---|
| $s$ | Number of desirable outputs |
| $h$ | Number of undesirable outputs |
| $g_{max}, g_{min}$ | Maximum and minimum normalized desirable output |
| $R_b$ | Weight of undesirable output |
| $b_{max}, b_{min}$ | Maximum and minimum normalized undesirable output |
| $Z$ | Objective function value (efficiency measure) |
| $y$ | Normalized desirable output |
| $b$ | Normalized undesirable output |
| $R_b$ | Undesirable output weight |
| $x_i$ | Inputs (Distance, Diesel Cost per km) |
| | |
| $y_i$ | Desirable output (Monthly Surplus Bagasse) |
| $b_i$ | Undesirable output (Transportation cost per ton) |
| **X** | Individual data values in the dataset |
| $\mu$ | Mean (average) of the dataset |
| $\sigma$ | Standard deviation of the dataset |
| **E** | Expected value (statistical average of the cubed deviations) |
| **X** | Values of the first variable in correlation |
| **Y** | Values of the second variable |
| $\bar{X}$ | Mean of variable |
| $\bar{Y}$ | Mean of variable |
| $\sum$ | Summation of all observations |
| **r** | Correlation coefficient indicating strength and direction of relationship |

**REFERENCES:**

[1] Jabbarzadeh, A., & Shamsi, M. (2025). Designing a resilient and sustainable multi-feedstock bioethanol supply chain: Integration of mathematical modeling and machine learning. *Applied Energy*, *377*, 123794. https://doi.org/10.1016/j.apenergy.2024.123794

[2] Bukhtoyarov, V., Tynchenko, V., Bashmur, K., Kolenchukov, O., Kukartsev, V., & Malashin, I. (2024). Fuzzy neural network applications in biomass gasification and pyrolysis for biofuel production: a review. *Energies*, *18*(1), 16. https://doi.org/10.3390/en18010016

[3] Malashin, I. P., Martysyuk, D., Nelyub, V., Borodulin, A., Gantimurov, A., & Tynchenko, V. (2026). Multi-objective optimisation in biofuel and bioenergy systems: a comprehensive review. *International Journal of Sustainable Engineering*, *19*(1), 2619227. https://doi.org/10.1080/19397038.2026.2619227

[4] Tehrani, M. S., Noori, S., & Dehghani, E. (2025). A Two-Stage Approach to Enhancing Biofuel Supply Chains Through Predictive and Optimization Analytics. *Supply Chain Analytics*, 100155. https://doi.org/10.1016/j.sca.2025.100155

[5] Hiloidhari, M., Sharno, M. A., Baruah, D. C., & Bezbaruah, A. N. (2023). Green and sustainable biomass supply chain for environmental, social and economic benefits. *Biomass and Bioenergy*, *175*, 106893. https://doi.org/10.1016/j.biombioe.2023.106893

[6] Huang, X., Ji, L., Xie, Y., & Luo, Z. (2025). Robust optimization of regional biomass supply chain system design and operation with data-driven uncertainties. *Food and Bioproducts Processing*, *149*, 176-189. https://doi.org/10.1016/j.fbp.2024.11.021

[7] Alinejad, S., Alimohammadlou, M., Abbasi, A., & Mirghaderi, S. H. (2024). Smart-circular strategies for managing biomass resource challenges: A novel approach using circular intuitionistic fuzzy methods. *Energy Conversion and Management*, *314*, 118690. https://doi.org/10.1016/j.enconman.2024.118690

[8] Wang, D., Ning, Y., Xiang, C., & Chen, A. (2024). A two-stage network framework for topology optimization incorporating deep learning and physical information. *Engineering Applications of Artificial Intelligence*, *133*, 108185. https://doi.org/10.1016/j.engappai.2024.108185

[9] Abandansari, S. A. M., Ghasemi, P., Attar, A., Goodarzian, F., & Ahmadi, S. A. (2025). A hybrid biomass supply chain optimization approach using sequential adaptive fuzzy inference learning: Insights from a cogeneration system in Europe. *Results in Engineering*, *26*, 105261. https://doi.org/10.1016/j.rineng.2025.105261

[10] Zhou, Z., Ma, Y., Pan, Y., & Zhu, Y. (2022). Data envelopment analysis models based on decentralized decision making. *Computers & Industrial Engineering*, *170*, 108318. https://doi.org/10.1016/j.cie.2022.108318

[11] Guerrero, N. M., Aparicio, J., & Valero-Carreras, D. (2022). Combining data envelopment analysis and machine learning. *Mathematics*, *10*(6), 909. https://doi.org/10.3390/math10060909

[12] Hahn, G. J., Brandenburg, M., & Becker, J. (2021). Valuing supply chain performance within and across manufacturing industries: A DEA-based approach. *International Journal of Production Economics*, *240*, 108203. https://doi.org/10.1016/j.ijpe.2021.108203

[13] Hong, J. D. (2023). Application of transformed two-stage network dea to strategic design of biofuel supply chain network. *Journal of Systems Science and Systems Engineering*, *32*(2), 129-151. https://doi.org/10.1007/s11518-023-5559-7

[14] Razmi, S., Mirzaee, H., Kumar, G., & Sowlati, T. (2025). Applications of Machine Learning for Decision Support in Biomass Supply Chains: A Systematic Review. *Computers & Chemical Engineering*, 109451. https://doi.org/10.1016/j.compchemeng.2025.109451

[15] Singh, A. P., & Yadav, S. P. (2021, December). Performance evaluation of DMUs using hybrid fuzzy multi-objective data envelopment analysis. In *International Conference on Nonlinear Applied Analysis and Optimization* (pp. 329-343). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-0597-3_23

[16] Bolos, V. J., Benitez, R., & Coll-Serrano, V. (2025). Conventional and Fuzzy Data Envelopment Analysis with deaR. *arXiv preprint arXiv:2506.03766*. https://doi.org/10.48550/arXiv.2506.03766

[17] Habibi, F., Chakrabortty, R. K., & Abbasi, A. (2023). Evaluating supply chain network resilience considering disruption propagation. *Computers & Industrial Engineering*, *183*, 109531. https://doi.org/10.1016/j.cie.2023.109531

[18] Gital, Y., & Bilgen, B. (2024). Biomass supply chain network design under uncertainty, risk and resilience: A systematic literature review. *Computers & Industrial Engineering*, *193*, 110270. https://doi.org/10.1016/j.cie.2024.110270

[19] Peng, W., & Sadaghiani, O. K. (2023). An analytical review on the utilization of machine learning in the biomass raw materials, their evaluation, storage, and transportation. *Archives of Computational Methods in Engineering*, *30*(8), 4711-4732. https://doi.org/10.1007/s11831-023-09950-9

[20] Peng, W., & Karimi Sadaghiani, O. (2023). A review on the applications of machine learning and deep learning in agriculture section for the production of crop biomass raw materials. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *45*(3), 9178-9201. https://doi.org/10.1080/15567036.2023.2232322

[21] Hai, T., Alsharif, S., Aziz, K. H. H., Dhahad, H. A., & Singh, P. K. (2023). Deep learning optimization of a biomass and biofuel-driven energy system with energy storage option for electricity, cooling, and desalinated water. *Fuel*, *334*, 126024. https://doi.org/10.1016/j.fuel.2022.126024
[22] Wang, J. (2024). Machine Learning Applications in Biomass Supply Chain Management and Optimization. *BioResources*, *19*(4). *DOI: 10.15376/biores.19.4.6961-6963*
[23] Han, C., & Zhang, Q. (2021). Optimization of supply chain efficiency management based on machine learning and neural network. *Neural Computing and Applications*, *33*(5), 1419-1433. https://doi.org/10.1007/s00521-020-05023-1
[24] Zhang, Z., Xiao, Y., & Niu, H. (2022). DEA and machine learning for performance prediction. *Mathematics*, *10*(10), 1776. https://doi.org/10.3390/math10101776
[25] Khedr, A. M. (2024). Enhancing supply chain management with deep learning and machine learning techniques: A review. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(4), 100379. https://doi.org/10.1016/j.joitmc.2024.100379
[26] Garg, R., Rajput, P., Vibhandik, J., Ali, A., & Abrar, I. (2025). Advances in ai-driven biomass processing: a review of conversion technologies, optimization strategies, and smart energy integration. *ACS omega*, *10*(42), 49300-49320. https://doi.org/10.1021/acsomega.5c05427
[27] Department of Co-operation, Marketing and Textiles, Maharashtra. Available at: https://mahasahakar.maharashtra.gov.in/en/sugar-commissionerate/
[28] Martin Matis, PhD. Cyril Sadovsky (2025), AI-Based Table Extraction from Technical Documents, DOI:10.13140/RG.2.2.25764.74885,
[29] Lei, S., & Li, Y. (2023). English machine translation system based on neural network algorithm. Procedia Computer Science, 228, 409-420. https://doi.org/10.1016/j.procs.2023.11.047
[30] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big data analytics, 1(1), 9. https://link.springer.com/article/10.1186/s41044-016-0014-0
[31] Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. Knowledge-Based Systems, 160, 104-118. https://doi.org/10.1016/j.knosys.2018.06.012
[32] Delaporte, G., Cladière, M., & Camel, V. (2019). Missing value imputation and data cleaning in untargeted food chemical safety assessment by LC-HRMS. Chemometrics and Intelligent Laboratory Systems, 188, 54-62. https://doi.org/10.1016/j.chemolab.2019.03.005
[33] Rangnekar, D. V. (1988). Integration of sugarcane and milk production in Western India Available at: https://www.fao.org/4/s8850e/S8850E17.htm
[34] Lintu, L. (1978). Panels, paper and paperboard from agricultural residues Available at: https://www.fao.org/4/L2015E/l2015e03.htm
[35] Misra, V., Shrivastava, A. K., Solomon, S., Ansari, M. I., & Shukla, S. P. (2016). Impact of crushing capacity of sugarcane in India. *Am.-Eurasian J. Agric. Environ. Sci*, *16*, 28-32 10.5829/idosi.aejaes.2016.16.1.12835
[36] Punnathanam, V., & Shastri, Y. (2022). Optimization-based design for lignocellulosic ethanol production: a case study of the state of Maharashtra, India. *Clean Technologies and Environmental Policy*, *24*(3), 863-886.DOI: https://doi.org/10.1007/s10098-021-02227-4
[37] Howitt, C., & Ghaffariyan, M. R. (2025). Truck configurations for efficient transportation of forest biomass: a short literature review. *Silva Balcanica*, *26*(1), 19-27 DOI: https://doi.org/10.3897/silvabalcanica.26.e138435
[38] Gilani, H., Sahebi, H., & Oliveira, F. (2020). Sustainable sugarcane-to-bioethanol supply chain network design: A robust possibilistic programming model. *Applied Energy*, *278*, 115653 https://www.sciencedirect.com/science/article/abs/pii/S030626192031151X
[39] Dudáš, A. (2024). Graphical representation of data prediction potential: correlation graphs and correlation chains. *The Visual Computer*, *40*(10), 6969-6982. https://link.springer.com/article/10.1007/s00371-023-03240-y
[40] Guo, H., Zhang, W., Ni, C., Cai, Z., Chen, S., & Huang, X. (2020). Heat map visualization for electrocardiogram data analysis. *BMC Cardiovascular Disorders*, *20*(1), 277.
[41] Gilani, H., Sahebi, H., & Oliveira, F. (2020). Sustainable sugarcane-to-bioethanol supply chain network design: A robust possibilistic programming model. *Applied Energy*, *278*, 115653 https://www.sciencedirect.com/science/article/abs/pii/S030626192031151X

[42] Paudel, S., Pudasaini, A., Shrestha, R. K., & Kharel, E. (2023). Compressive strength of concrete material using machine learning techniques. *Cleaner Engineering and Technology*, *15*, 100661 https://www.sciencedirect.com/science/article/pii/S0957417420303457

[43] Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of property research, 38(1), 48-70 https://doi.org/10.1080/09599916.2020.1832558

[44] Garcia-Gutierrez, J., Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C. (2014). A Comparative Study of Machine Learning Regression Methods on LiDAR Data: A Case Study. In: Herrero, Á., et al. International Joint Conference SOCO'13-CISIS'13-ICEUTE'13. Advances in Intelligent Systems and Computing, vol 239. Springer, Cham. https://doi.org/10.1007/978-3-319-01854-6_26

[45] Xiaowei Jin, Wenbin Guo, Tianyu Shi, Zuocheng Wang, Machine learning-based prediction and optimization of biomass pellet strength for sustainable bioenergy production, Industrial Crops and Products, Volume 234, 2025, 121498, ISSN 0926-6690, https://doi.org/10.1016/j.indcrop.2025.121498

[46] Hemmati, H., Kazemzadeh, R.B., Nikbakhsh, E. *et al.* Resilient and Sustainable Supplier Performance Evaluation in Uncertain Conditions: A Novel Comprehensive Approach with Fuzzy Data Envelopment Analysis. *J. Syst. Sci. Syst. Eng.* (2025). https://doi.org/10.1007/s11518-025-5655-y

[47] Bationo F., Bayala B. A., Zaida J. T.(2026). Optimization Of The Concrete Mixer Manufacturing Process: An Integrated Approach Combining Mathematical Modeling And Context-Appropriate Technical Solutions For Burkina Faso. International Journal of Engineering Sciences & Research Technology, 15(4), 27-42. https://doi.org/10.64149/j.ijesrt.15.5.27-42

[48] Tavassoli, M. Measuring fair efficiency decomposition in network DEA model under uncertainty: modeling and computational aspects for sustainable supply chain performance assessment. *Oper Res Int J* **25**, 64 (2025). https://doi.org/10.1007/s12351-025-00936-y

[49] Hassan M., (2026). Rheological Properties and Optimization for Different Types of Concrete: A Review. International Journal of Engineering Sciences & Research Technology, 15(5), 83-93 https://doi.org/10.64149/j.ijesrt.15.5.83-93

[50] Masoud Vaseei, Maryam Daneshmand-Mehr, Morteza Bazrafshan, Armin Ghane Kanafi, A network data envelopment analysis to evaluate the performance of a sustainable supply chain using bootstrap simulation, Journal of Engineering Research, Volume 12, Issue 4, 2024, Pages 904-915, ISSN 2307-