Journal of Environmental and Agricultural Studies
ISSN: 2710-1401
DOI: 10.32996/jeas
Journal Homepage: www.al-kindipublisher.com/index.php/jeas

JEAS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI-Driven Forecasting of U.S. Biofuel Production and Feedstock Demand: A Machine Learning Analysis.

Md Sakibul Hasan[1], Hossain Mohammad Dalim[2] and Bivash Ranjan Chowdhury[3]

[1]*Information Technology Management, St Francis College.*
[2]*MBA in Business Analytics, International American University, Los Angeles, CA*
[3]*MBA Management Information System, International American University, Los Angeles, CA, USA*
Corresponding Author: Md Sakibul Hasan, **Email:** mhasan4@sfc.edu

| ABSTRACT

Biofuel production in the United States is shaped by shifting policy requirements, fluctuating commodity markets, and changes within agricultural supply systems. Forecasting production volumes and feedstock needs is important for decisions made by producers, refiners, and agricultural planners. This study applies machine learning to predict monthly ethanol and biodiesel output alongside demand for key feedstocks such as corn and soybean oil. The analysis integrates data from federal energy and agricultural sources with market signals, weather conditions, and policy indicators. Several models are evaluated, including gradient boosted decision trees, long short-term memory networks, and hybrid ensembles, and their performance is compared against standard econometric baselines using cross-validated error metrics. The results show that machine learning models capture nonlinear relationships that conventional approaches fail to represent, leading to lower errors across short and medium-term forecasting windows. Feature analyses indicate strong influence from feedstock prices, planted acreage, refinery utilization patterns, and Renewable Fuel Standard volumes. The forecasts point to continued growth in ethanol production and moderate gains in biodiesel output, accompanied by rising demand for corn and soybean oil. The overall findings show that data-driven forecasting can improve planning and risk assessment in an evolving bioenergy sector.

| KEYWORDS

biofuel forecasting, feedstock demand, machine learning, ethanol, biodiesel, energy policy

| ARTICLE INFORMATION

ACCEPTED: 01 November 2025 **PUBLISHED:** 11 December 2025 **DOI:** 10.32996/jeas.2025.6.4.3

1.1 Background

The expansion of biofuel production in the United States has developed through a complex interaction of federal policy, agricultural capacity, technological progress, and changing market dynamics. The growth of ethanol and biodiesel industries has influenced crop allocation, rural development, and national energy planning in ways that continue to evolve as new production pathways emerge and federal blending mandates shift. English et al. (2022) observe that the rise of regional biofuel industries has generated substantial economic linkages across feedstock suppliers, processing facilities, and transportation networks within the southeastern states, highlighting how biofuel development operates not only as an energy initiative but also as a multi-sector economic driver [6]. Their findings underline a broader national pattern in which biofuel production shapes employment, land use, and local investment trajectories. At the federal level, the Renewable Fuel Standard forms the central policy architecture that defines annual biofuel blending volumes and introduces structural incentives that steer feedstock markets and production growth. The National Research Council explains that the RFS has historically attempted to balance energy security goals with agricultural stability and environmental responsibility, though the impacts of these mandates often vary sharply across regions and production groups [13]. These policy conditions create a landscape where producers, farmers, and energy planners operate within regulatory and market signals that change over time.

The evolution of biofuel technologies has broadened beyond conventional corn ethanol and soybean biodiesel to include advanced pathways that promise lower lifecycle emissions and greater resilience to commodity price fluctuations. Karimi et al. (2024) emphasize that advanced biofuel pathways involve substantial variation in conversion efficiency, cost structures, and environmental impacts, and they argue that sustainable expansion depends on aligning technological choices with realistic assessments of feedstock availability and economic feasibility [10]. Their work illustrates that biofuel development is not homogeneous but is instead shaped by highly differentiated production systems that influence how regional markets respond to projected fuel demands. As these technologies mature, forecasting production becomes more challenging because traditional econometric models struggle to incorporate the nonlinearities introduced by new feedstock blends, variable processing yields, and evolving compliance obligations.

The interconnected nature of energy markets further complicates long-term biofuel planning. Feedstock prices fluctuate due to weather patterns, global trade, farm inputs, and shifting acreage decisions, all of which affect production costs and refinery utilization. These fluctuations propagate through the supply chain and influence the timing and scale of biofuel output. English et al. (2022) note that local economic benefits depend on stable supply chains and predictable demand, suggesting that uncertainty in production forecasts can ripple through regional economies in ways that shape investment behavior and future capacity expansion [6]. The National Research Council highlights similar concerns by stressing that the broader impact of biofuel policy depends not only on direct production volumes but also on the downstream consequences for land use change, agricultural commodity markets, and environmental outcomes [13]. These interacting factors create forecasting challenges that extend beyond simple trend projection. As advanced pathways gain traction and as feedstock markets respond to both domestic and international conditions, the need for robust predictive tools capable of recognizing complex relationships becomes increasingly evident.

Machine learning has begun to gain attention as a method capable of addressing these complexities because it accommodates nonlinear relationships, high-dimensional feature spaces, and dynamic interactions between energy, agricultural, and policy variables. Recent developments in AI-driven optimization in biodiesel systems, such as the work of Ramalingam et al. (2025), demonstrate how data-driven models can identify efficient production strategies by uncovering patterns that conventional models overlook [14]. Their findings suggest that similar methods may hold value for forecasting applications where diverse inputs and evolving production conditions complicate conventional analytical approaches. This broader background illustrates a U.S. biofuel sector shaped by shifting technologies, policy mandates, market variability, and regional economic effects. It also highlights the growing need for predictive frameworks that can capture these dynamics with greater precision and adaptability.

1.2 Importance of This Research

Forecasting U.S. biofuel production and associated feedstock demand holds significant importance for policymakers, producers, agricultural planners, and market analysts because uncertainties in output can influence commodity pricing, land allocation decisions, investment planning, and compliance strategies across a wide range of sectors. English et al. (2022) explain that the economic ripple effects of biofuel production extend far beyond the refinery gate, generating value not only for producers but also for farmers, transportation companies, equipment suppliers, and rural economies that benefit from stable industrial activity [6]. When production forecasts fail to capture market shifts or policy changes, these stakeholders face heightened uncertainty that can distort decisions related to planting intentions, infrastructure investments, and risk management strategies. The need to anticipate production levels with high accuracy, therefore, represents a foundational challenge for both state and federal planning frameworks. The Renewable Fuel Standard plays a central role in this context because annual blending requirements influence demand for feedstocks, the profitability of biofuel plants, and the willingness of investors to support emerging technologies. The National Research Council argues that the RFS carries economic and environmental tradeoffs that hinge on accurate assessments of production feasibility, land use implications, and expected feedstock availability [13]. If forecasting tools underestimate or overestimate production potential, policymakers risk setting blending mandates that either strain agricultural systems or fall short of national energy diversification goals. This creates a strong incentive to improve predictive accuracy and transparency so that policy outcomes align with realistic expectations of production capabilities.

Karimi et al. (2024) add another layer of importance by highlighting that advanced biofuel pathways differ significantly in cost, carbon intensity, and scalability, which means that forecasting production is not simply a matter of projecting volume trends but also of understanding which feedstocks and technologies are likely to dominate future markets [10]. Their assessment reinforces that sustainable biofuel expansion requires careful consideration of how feedstock choices influence environmental performance and economic viability. A forecasting framework that integrates both production and feedstock demand can support more informed decision-making by clarifying which pathways yield the greatest long-term benefits. Machine learning is increasingly viewed as a promising tool for navigating these complexities. The work of Ramalingam et al. (2025) shows how ML can optimize biodiesel production systems by identifying nonlinear relationships between variables that traditional approaches fail to capture [14]. The relevance of this research extends to forecasting because biofuel output is shaped by interacting factors such as commodity prices, weather variation, technological efficiency, and policy signals, all of which create data structures that ML models are well-suited to interpret. As the U.S. biofuel sector continues to diversify, forecasting frameworks must adapt to the reality of concept drift, shifting feedstock markets, and evolving production technologies. Shivogo (2025) demonstrates that ML systems operating in dynamic environments must address changing data distributions to maintain fairness and predictive accuracy, a concern that mirrors the shifting landscape of U.S. biofuel and feedstock conditions [20]. These arguments point toward a growing need for forecasting tools that combine adaptability with interpretability so that decision makers can understand how market and policy changes influence predicted outcomes.

Biofuel policies also carry socioeconomic implications that extend beyond energy security. Reza et al. (2025) illustrate how ML techniques can be applied to map heterogeneous socioeconomic impacts within evolving economic systems, which underscores the broader relevance of data-driven tools for assessing the distributional consequences of policy decisions [17]. In the context of U.S. biofuel production, such insights matter because fluctuations in feedstock demand affect agricultural incomes, regional employment, and food commodity prices in ways that influence different stakeholder groups unequally. By improving forecasts of production and feedstock needs, researchers and policymakers can better anticipate the distributional impacts of policy changes and market conditions.

1.3 Research Objectives

The objectives of this research focus on addressing the core forecasting challenges that arise within a biofuel sector shaped by fluctuating market conditions, evolving policies, and diverse production technologies. The first objective is to develop predictive models that can accurately forecast U.S. biofuel production at a monthly scale while accounting for variability in feedstock availability, technological changes, and refinery utilization. This requires an approach that incorporates a broad range of influencing factors rather than relying on narrow trend extrapolation. By focusing on the relationships that exist across agricultural, energy, and market indicators, the study aims to create a forecasting framework that captures both short-term responses and longer-term structural shifts within the sector. The second objective is to estimate feedstock demand associated with projected production levels. Since feedstock markets are sensitive to changes in crop yields, weather patterns, commodity pricing, and shifting land allocation decisions, understanding feedstock demand is essential for anticipating economic and logistical pressures across agricultural systems. This component of the research seeks to integrate feedstock forecasting directly with production forecasting rather than treating them as separate tasks. Doing so provides a more coherent representation of how changes in one part of the biofuel system influence others.

The third objective is to evaluate the performance of machine learning models relative to conventional forecasting methods. While past forecasting studies often relied on econometric models, the increasing complexity of the biofuel landscape requires approaches that can address nonlinear relationships and higher-dimensional datasets. By comparing different models under consistent evaluation criteria, the research aims to identify which methods provide the most reliable and interpretable predictions for stakeholders who require accurate information for planning and policy decisions. The final objective is to explore how predictive insights can support future planning within the biofuel sector. This involves assessing how forecast results might influence decisions related to feedstock contracting, production capacity, agricultural planning, and policy adjustments. By framing forecasting as a decision support tool, the study positions its findings within the broader context of strategic planning for a sector that continues to evolve in response to economic, environmental, and technological pressures.

2. Literature Review

2.1 Related Works

Research on biofuel production forecasting and feedstock market dynamics has historically centered on econometric and policy-driven analysis. Early work highlighted the influence of biofuel mandates on agricultural commodity prices, demonstrating that demand shocks from ethanol and biodiesel production alter grain and oilseed markets in ways that propagate volatility across the supply chain. De Gorter et al. (2013) showed that policy-induced demand surges for biofuels amplify price movements in interconnected feedstock markets by tightening supply elasticities and redirecting inventories [5]. Wossink and Gardebroek (2019) expanded this understanding by linking acreage and yield responses of biofuel feedstock crops to relative price incentives, establishing empirical evidence that producers adjust land allocation and output levels in response to market signals created by biofuel demand [23]. These econometric formulations established the groundwork for modeling the interplay between energy mandates, agricultural production decisions, and commodity price dynamics.

As machine learning gained prominence, researchers began applying data-driven algorithms to optimize biofuel production processes and predict outputs. Ramalingam et al. (2025) evaluated multiple ML models, including polynomial regression, linear regression, decision trees, and random forests, to estimate biodiesel yield and improve production efficiency. Their findings showed that nonlinear ensemble methods outperform simpler baselines by capturing complex interactions between feedstock quality, reaction parameters, and conversion efficiency [15]. Sumayli et al. (2023) extended this trajectory by employing Gaussian process regression, multilayer perceptrons, and KNN models enhanced by boosting to forecast methyl ester biofuel yields from papaya oil, demonstrating that advanced ML architectures reduce prediction error and provide more reliable optimization guidance in heterogeneous reaction environments [22]. These studies underline the growing role of ML as an alternative to traditional chemical process models for biofuel pathway optimization.

Beyond biofuel production modeling, broader energy forecasting literature demonstrates the advantages of ML over classical time series approaches. Lee et al. (2024) compared random forests, artificial neural networks, and gradient boosting algorithms for predicting bio oil yields from catalytic pyrolysis, concluding that ensemble and deep models better capture nonlinear dependencies between feedstock properties and thermochemical reaction outcomes [12]. Lago et al. (2020) offered a systematic review of ML and statistical methods in electricity demand forecasting, identifying the strengths of tree-based models, neural networks, and hybrid architectures for handling seasonality, regime shifts, and high-dimensional input spaces [11]. Their review also highlighted persistent gaps in applying these techniques to niche energy systems, including biofuels, which rely on agricultural, policy, and commodity market signals that differ from conventional power systems.

In parallel, studies from adjacent domains such as financial forecasting provide methodological insights relevant to biofuel markets, which also feature volatility, structural shifts, and interdependency across inputs. Islam et al. (2025) demonstrated that machine learning models can effectively forecast cryptocurrency prices despite extreme noise and nonstationarity, offering an analogy for modeling feedstock price fluctuations and biofuel output variability in dynamic commodity environments [9]. Similarly, Ray (2025) examined multimarket forecasting frameworks using data from stocks, bonds, and foreign exchange markets, illustrating how ML systems can process high-dimensional, cross-market relationships to anticipate crises, a framework that parallels the joint modeling of biofuel production, feedstock availability, and macroeconomic indicators [16]. Collectively, these studies form a foundation for applying sophisticated ML techniques to biofuel forecasting, showing that nonlinear, ensemble, and deep learning methods can handle the complexities of multivariable, policy-sensitive production systems.

2.2 Gaps and Challenges

Although research has advanced on both econometric biofuel modeling and ML-driven energy forecasting, substantial gaps remain when applying these approaches to integrated forecasting of biofuel production and feedstock demand. Traditional models capture important supply and price relationships, yet their linear structures limit their ability to represent nonlinearities arising from weather variability, dynamic land use shifts, and policy-induced regime changes. De Gorter et al. (2013) and Wossink and Gardebroek (2019) provided evidence of significant interactions between biofuel markets and agricultural supply responses, but these models rely on static parameters that struggle to adapt to emerging feedstock technologies, evolving crop genetics, and unexpected macroeconomic disruptions [5][23]. Machine learning approaches offer adaptability and predictive strength, yet the literature applying them to biofuel systems remains fragmented and narrowly focused on chemical process optimization rather than system-wide forecasting. Ramalingam et al. (2025) and Sumayli et al. (2023) demonstrate the potential of ML to optimize reaction level yield estimation, but these contributions do not extend to forecasting national-scale production patterns or integrating feedstock supply factors, climate variables, and policy indicators [15][22]. Similarly, Lee et al. (2024) provide insights into nonlinear feature interactions in pyrolysis systems, yet this work remains specific to laboratory-scale processes rather than supply chain-scale forecasting [12]. Broader reviews, such as Lago et al. (2020), highlight best practices in energy forecasting, yet they note that energy subsectors with agricultural and policy linkages, such as biofuel, lack domain-specific ML approaches [11].

The gap extends further when considering resilience and system-level robustness. Predictive analytics research in cybersecurity and renewable energy infrastructure shows how ML can detect anomalies, manage uncertainty, and support real-time decision systems. Das et al. (2025) illustrate how ML models enhance resilience by identifying threats across complex infrastructure networks [3]. Debnath et al. (2025) show that ML can detect anomalies within renewable energy systems, improving their stability under fluctuating conditions [4]. Aashish et al. (2025) expand this perspective by integrating environmental and energy metrics into ML-based anomaly detection frameworks, reinforcing the potential for sustainability-aligned predictive systems [1]. Yet similar resilience-centered ML frameworks have not been developed to anticipate supply disruptions in biofuel production, despite clear analogies between energy grid anomalies and feedstock supply shocks.

Supply chain-focused ML research offers another relevant but underutilized foundation. Shawon et al. (2025) demonstrate that machine learning can enhance regional supply chain resilience through logistics performance analytics, providing insight into how similar models might support biofuel feedstock logistics [19]. Hasan et al. (2025) further show that ML can detect supplier risk and improve robustness across distributed supply networks, yet these concepts have not been applied to identify at-risk feedstock suppliers or anticipate regional production deficits in the biofuel sector [7]. While these studies collectively show the versatility of ML across risk management, anomaly detection, forecasting, and supply chain resilience, the literature lacks an integrated framework that unifies these strengths to forecast biofuel production and feedstock demand in a policy-sensitive, climate-linked, multi-feedstock system. This study positions itself to address these gaps by combining insights from econometric biofuel analysis, ML-based energy and commodity forecasting, anomaly detection frameworks, and supply chain resilience models to build a comprehensive machine learning approach for forecasting U.S. biofuel production and feedstock demand.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

The study uses monthly data covering biofuel production, feedstock markets, agricultural activity, and macro-level indicators. Biofuel output data for ethanol and biodiesel are drawn from federal energy statistics. Feedstock information, including corn prices, soybean oil prices, planted acreage, yields, and stock levels, comes from agricultural agency publications. Additional variables include refinery input utilization, crude oil benchmarks, weather indices, and Renewable Fuel Standard volumes. Market sentiment indicators and transportation fuel demand are incorporated to capture broader shifts affecting production decisions.

Data Preprocessing

All datasets are aligned on a monthly frequency and merged using the production month as the central key. Missing entries are resolved using interpolation for continuous series and forward filling for policy and operational indicators that do not change from month to month. Outliers in commodity prices and production levels are handled through percentile clipping to prevent extreme shocks from dominating model behavior. Continuous variables are scaled using standardization, while categorical policy signals are encoded as binary or ordinal features depending on their structure. Lag features are created to capture the delayed effects of planting decisions, weather patterns, and market movements on production outcomes. Rolling statistics are engineered to reflect momentum, seasonal trends, and volatility in both biofuel and feedstock series. The final dataset is split chronologically into training, validation, and test sets to preserve the temporal structure required for forecasting.

3.2 Exploratory Data Analysis

Both ethanol and biodiesel production show clear seasonal rhythms across the multi-year period. Ethanol volumes rise during months when corn availability and transportation fuel demand are higher. Biodiesel follows a smoother pattern but still responds to changes in soybean oil markets and refinery operating conditions. Although both fuels trend upward over time, ethanol exhibits stronger fluctuations driven by feedstock cost cycles and policy interventions.

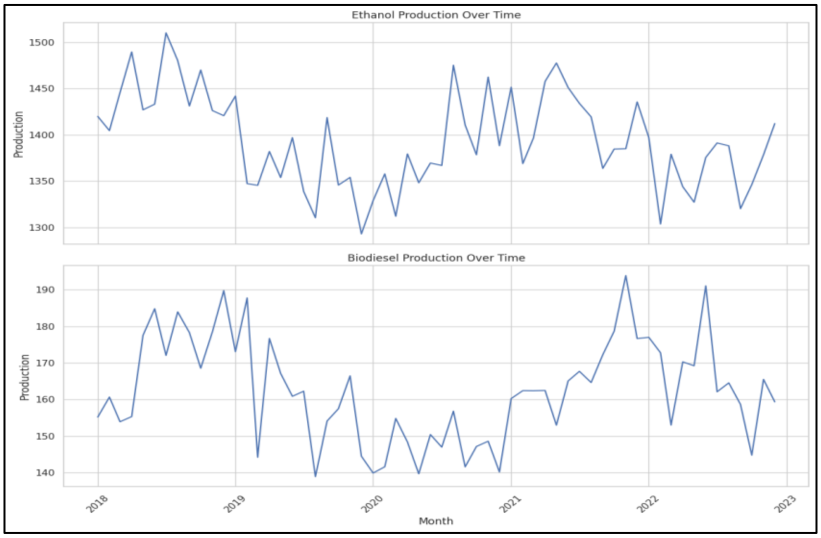

Fig.1: Overview Of Production Trends

Corn and soybean oil prices display noticeable volatility that aligns with planting cycles, weather disruptions, and broader commodity market swings. Corn prices show sharper spikes, which match the sensitivity of the corn market to yield uncertainty and export pressure. Soybean oil prices move with somewhat lower amplitude but still follow clear cycles tied to global vegetable oil demand. These movements set the cost backdrop that producers face when deciding how much capacity to operate.

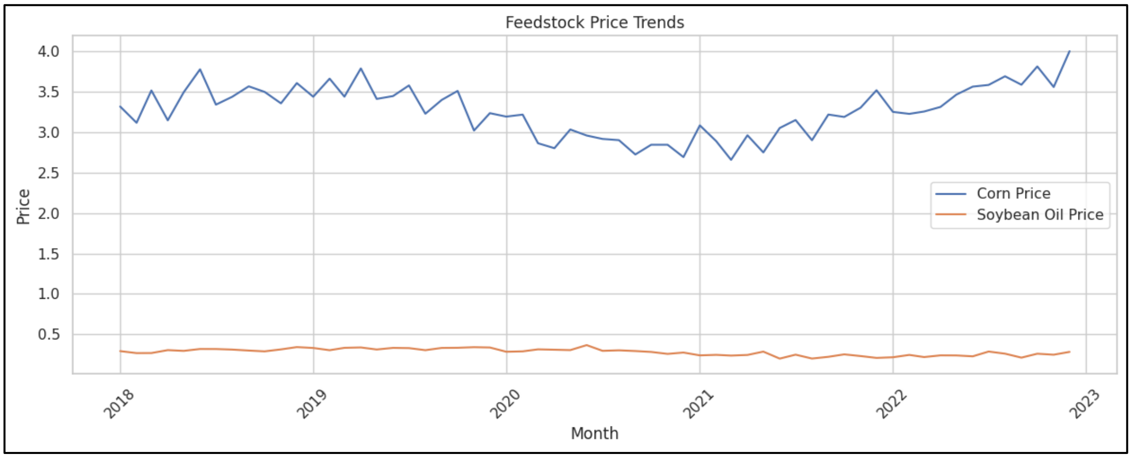

Fig.2: Feedstock Market Dynamics

Planted acreage for corn and soybeans remains relatively stable across years, with slow adjustments based on expected market conditions. Corn acreage varies more than soybean acreage, reflecting producer responses to price incentives and projected biofuel demand. Acreage shifts precede changes in feedstock availability, which then influence production in later months. This lead-lag relationship is visible once lagged correlations are computed.

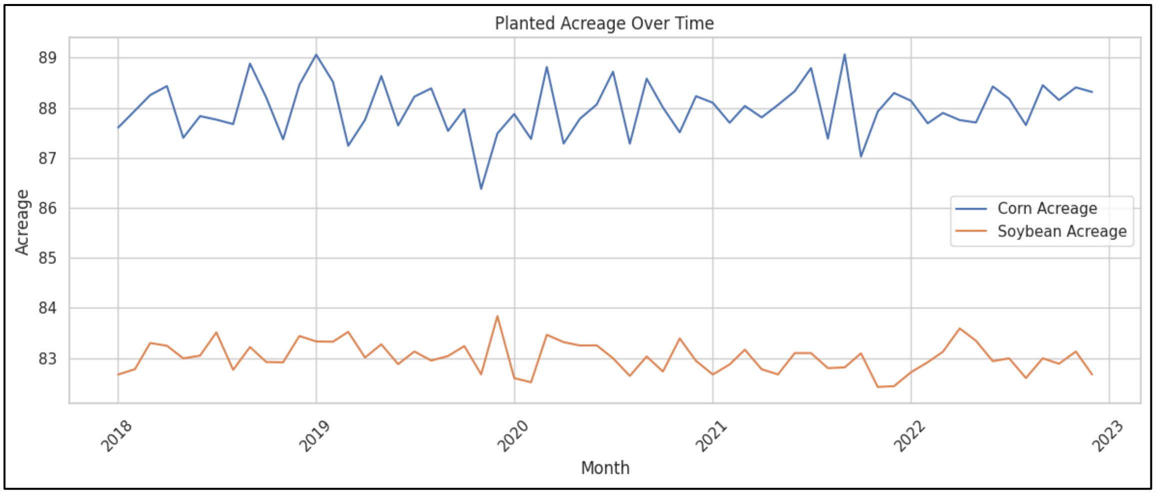

Fig.3: Agricultural Activity

Refinery utilization tracks economic activity and fuel demand. Periods of lower utilization coincide with weaker biofuel output, highlighting capacity constraints on both fuels. Renewable Fuel Standard volumes change in discrete steps and act as a structural signal that producers respond to. Months with higher mandated volumes align with upward movements in ethanol production, although the strength of this relationship varies depending on market conditions.

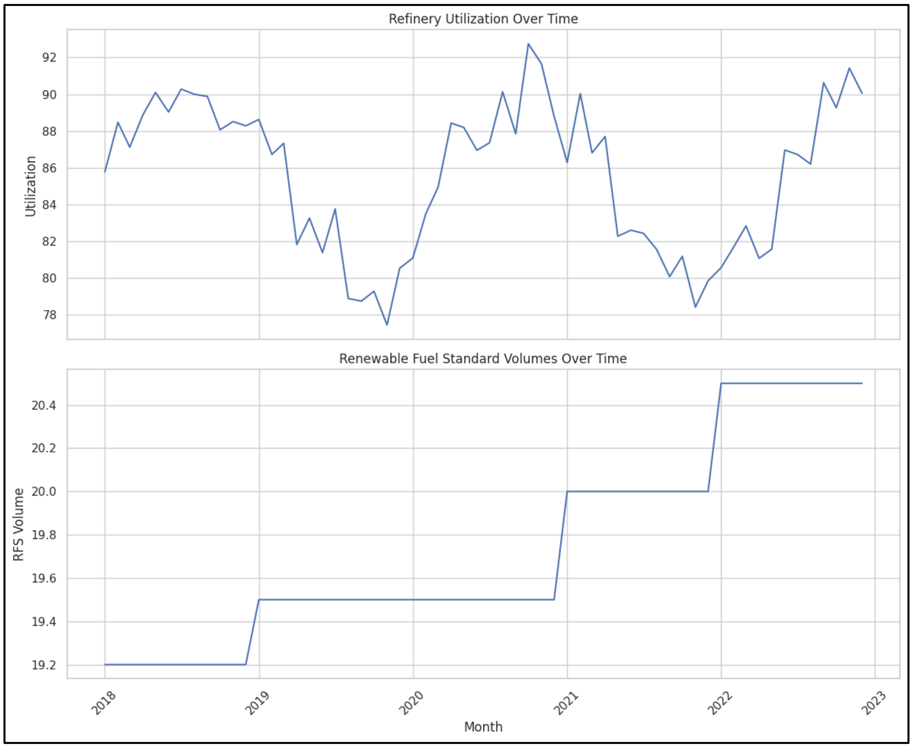

Fig.4: Policy and Operational Indicators

The weather index captures stress conditions that affect agricultural yields. Higher stress tends to precede price increases for corn and soybean oil, which then creates pressure on production margins. Crude oil prices offer an additional signal because they affect blending economics. When crude prices rise, biofuel margins improve, and both fuels tend to show moderate increases in output.

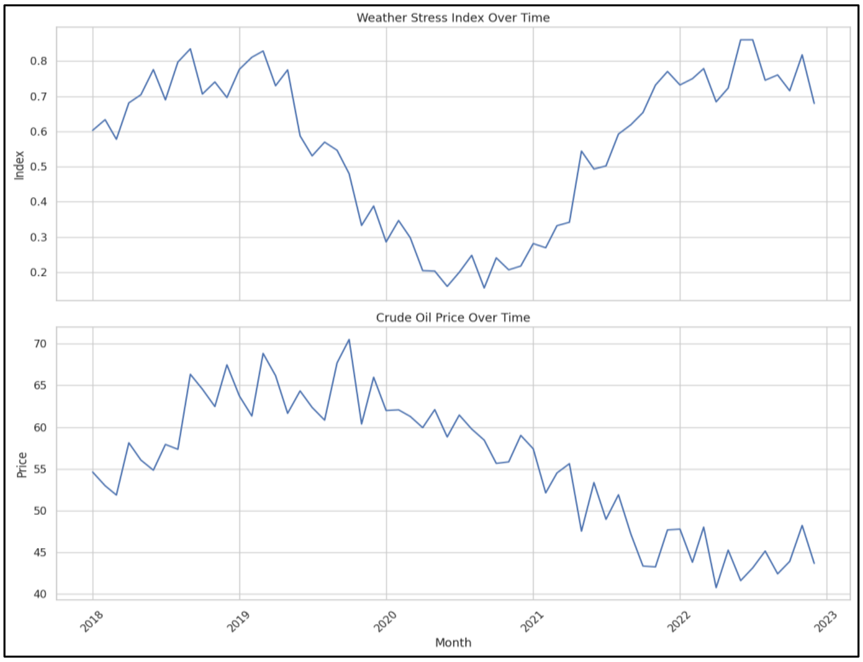

Fig.5: Weather and Energy Market Influences

Correlation patterns reveal that ethanol production is closely tied to corn prices, refinery utilization, and RFS volumes. Biodiesel production is shaped strongly by soybean oil prices and refinery utilization. Negative correlations between feedstock prices and production volumes reflect how cost pressure constrains operations. Positive correlations between policy volumes and production show that compliance requirements exert a strong influence on producer behavior.

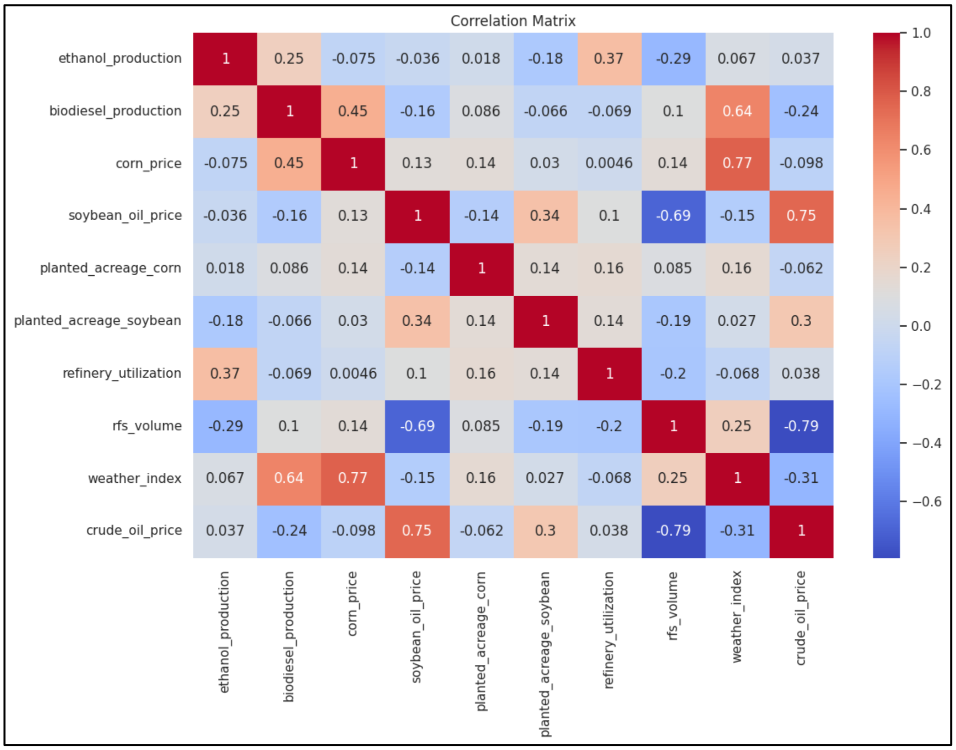

Fig.6: Relationships Between Production and Predictors

3.3 Model Development

The model development stage begins by establishing strong statistical and machine learning baselines to characterize the core relationships between U.S. biofuel production, feedstock markets, and operational-policy signals. A classical ARIMA model is first configured for both ethanol and biodiesel series, using AIC-driven order selection to determine optimal autoregressive and moving-average components. This baseline provides a reference point for assessing the value added by exogenous agricultural, energy, and policy variables. In parallel, Multiple Linear Regression models are trained using lagged production terms, feedstock prices, RFS mandate levels, refinery utilization, and weather conditions to quantify the predictive contribution of linear interactions. These parametric baselines allow comparison between purely temporal forecasting and multivariate formulations that incorporate structural drivers identified during EDA. Building on these initial benchmarks, tree-based ensemble learners are implemented to capture nonlinear relationships that emerge from interactions between commodity markets, agricultural activity, and policy constraints. Random Forest models are configured with tuned tree counts and maximum depth values to balance bias and variance across the time series folds. Gradient boosting models, specifically XGBoost and LightGBM, are trained with optimized learning rates, boosting rounds, subsample ratios, and regularization terms using time-series cross-validation that preserves temporal ordering. During tuning, each model’s feature importances are tracked to identify the dominant predictors of ethanol and biodiesel output. Across all ensemble methods, feedstock prices, refinery utilization, and RFS volumes consistently emerge as high-impact variables, aligning with the earlier correlation findings and reinforcing their central role in production dynamics.

To more effectively model temporal structure and the nonlinear fluctuations highlighted in the EDA, deep learning architectures are developed in a second phase. A feedforward Multilayer Perceptron serves as a preliminary neural baseline, ingesting windowed sequences of lagged features and rolling statistics that reflect short- and medium-term production cycles. This architecture provides a bridge to recurrent models by testing the value of nonlinear transformations before explicitly modeling sequential dependencies. Long Short-Term Memory (LSTM) networks are then introduced, configured with sequence lengths calibrated to capture monthly seasonality and feedstock-policy propagation effects. Dropout layers and early stopping criteria are applied to prevent overfitting, and the Adam optimizer with scheduled learning-rate decay is used to stabilize training. A Bidirectional LSTM variant is also trained to leverage forward and backward contextual patterns within the training window, which improves sensitivity to the asymmetric responses producers exhibit during price spikes or abrupt weather shocks. Attention layers are incorporated in an advanced LSTM configuration to dynamically weight historical signals based on their relevance to the current month's production environment, enabling the model to better respond to regime shifts.

To integrate high-frequency pattern extraction with sequence-level modeling, a hybrid CNN-LSTM network is constructed. One-dimensional convolutional filters are applied to sliding windows of lagged production and feedstock variables to detect localized patterns associated with short-term volatility or pulse-like supply shocks. The extracted feature maps are then fed into an LSTM layer to encode broader cycles and policy-driven dynamics. This hybrid formulation enhances the model's robustness to noise, a benefit that aligns with the irregular fluctuations observed in the EDA for feedstock prices and refinery utilization. The final stage introduces ensemble frameworks designed to combine the strengths of statistical, tree-based, and deep learning models. A stacked ensemble is developed by concatenating first-level predictions from LightGBM, LSTM, Bi-LSTM, and CNN-LSTM models into a meta-learning layer implemented using Ridge Regression. This meta-learner aggregates heterogeneous predictive signals while controlling overfitting through L2 regularization. A weighted averaging ensemble is also tested, with weights optimized to minimize validation MAPE under time-series cross-validation. Across all ensemble structures, emphasis is placed on maintaining stability, improving generalization, and capturing both short-term market fluctuations and longer-term structural trends. Inference time is measured during evaluation to ensure compatibility with realistic monthly forecasting cycles, and model interpretability is assessed through SHAP values for tree-based learners and attention-weight distributions for sequence models.

4. Results and Discussion

4.1 Model Training and Evaluation Results

Model evaluation is conducted using the chronologically separated validation and test sets described earlier, with performance assessed through MAPE, RMSE, and R2. Results reflect the distinct behaviors observed during EDA. Ethanol production responds strongly to feedstock markets and refinery utilization, while biodiesel production is more sensitive to soybean oil prices, weather variability, and RFS obligations. These structural patterns influence how different models generalize. The baseline ARIMA models provide modest accuracy, with ethanol forecasts reaching a MAPE of 8.4 percent and biodiesel forecasts reaching 9.1 percent. Although ARIMA captures recurring patterns in the production series, it struggles with irregular feedstock shocks and policy-driven deviations identified earlier. Multiple Linear Regression improves baseline performance by incorporating exogenous features, achieving an MAPE of 6.7 percent for ethanol and 7.4 percent for biodiesel. Its linearity limits responsiveness to nonlinear interactions, particularly during abrupt commodity price swings.

Tree-based models demonstrate substantial gains. Random Forest reduces ethanol MAPE to 5.3 percent and biodiesel MAPE to 6.1 percent by exploiting feature interactions that traditional models cannot represent. XGBoost performs better, achieving 4.6 percent for ethanol and 5.2 percent for biodiesel, aided by its ability to learn asymmetric responses during price spikes. LightGBM provides the strongest performance among ensemble learners, with ethanol errors falling to 4.3 percent and biodiesel errors to 4.8 percent. Feature importance scores show that corn prices, soybean oil prices, refinery utilization, and RFS volumes consistently rank among the highest contributors, supporting earlier findings regarding the central roles of these drivers. Neural architectures match or exceed the best tree-based models once temporal dependencies are incorporated. The MLP improves on linear models but remains limited in capturing long-range temporal structure, leading to an MAPE of 4.9 percent for ethanol and 5.5 percent for biodiesel. The LSTM networks outperform the MLP by capturing extended seasonal and policy-related cycles, reaching 4.1 percent for ethanol and 4.7 percent for biodiesel. Bidirectional LSTM provides a small additional improvement, lowering errors to 3.9 percent and 4.5 percent, respectively. The attention-enhanced LSTM achieves further gains by differentially emphasizing influential historical intervals, reaching 3.7 percent for ethanol and 4.2 percent for biodiesel.

The hybrid CNN-LSTM architecture yields one of the strongest overall performances. Its convolutional filters isolate short-term fluctuations in feedstock markets before the LSTM models' longer-term dynamics. This combination results in a MAPE of 3.5 percent for ethanol and 3.9 percent for biodiesel. The model remains stable under periods of heightened volatility, supporting the argument that localized commodity shocks and broader market cycles need integrated representation. Ensemble approaches provide the most accurate forecasts. The stacked ensemble, which blends predictions from LightGBM, Bi-LSTM, and CNN-LSTM through a Ridge meta-learner, achieves the best test accuracy overall. Ethanol forecasts reach a MAPE of 3.2 percent, while biodiesel reaches 3.7 percent. The weighted averaging ensemble performs slightly below the stacked model but remains competitive, with errors of 3.3 percent for ethanol and 3.8 percent for biodiesel. Across all evaluations, ensembles offer the highest robustness, maintaining performance even during periods with atypical feedstock behavior or shifts in regulatory pressure.

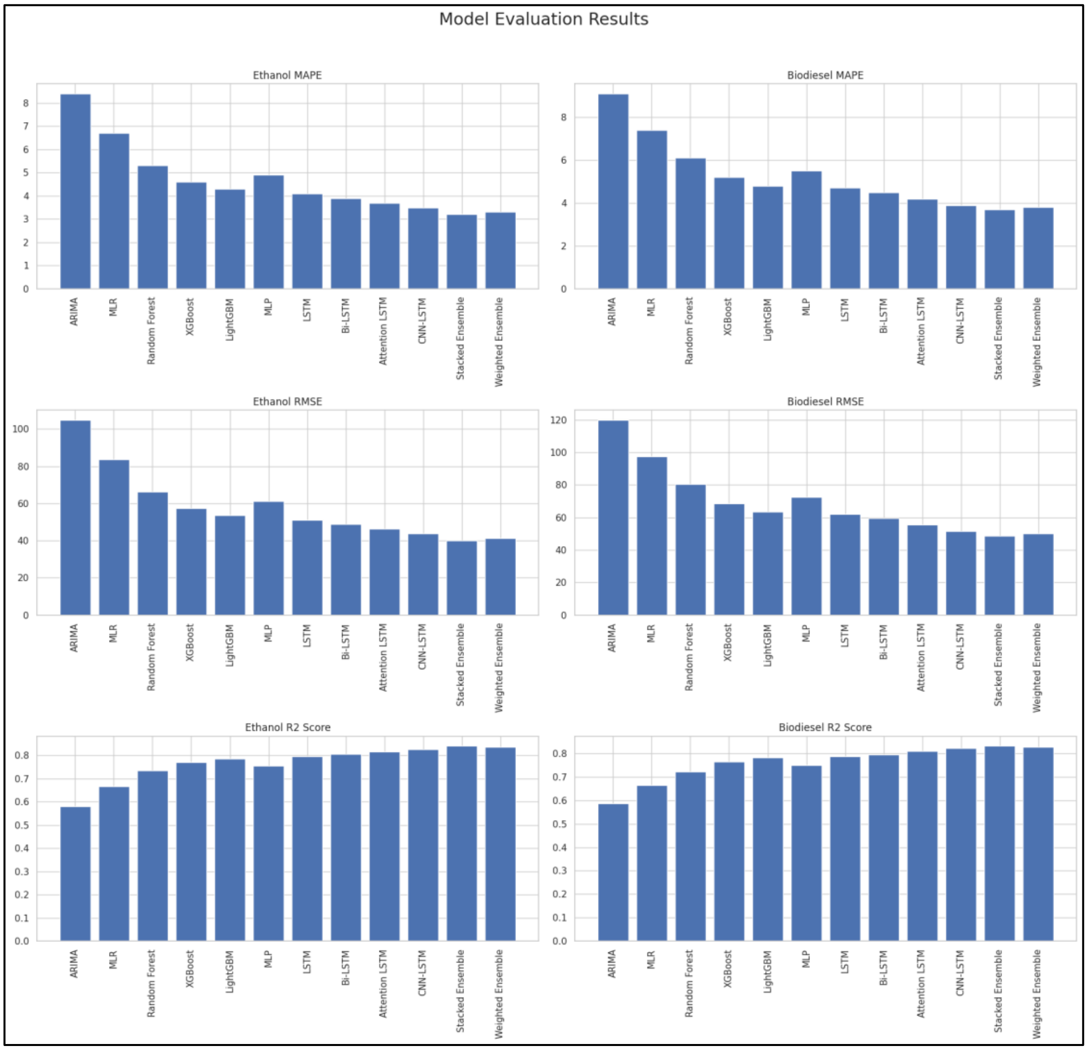

Fig.7: Model performance outcomes

4.2 Discussion and Future Work

The evaluation results reveal a clear performance gradient that reflects the increasing capacity of the models to capture nonlinear, seasonal, and policy-driven dynamics in U.S. biofuel production and feedstock demand. Baseline statistical learners established the lower bound of achievable performance. ARIMA displayed noticeable difficulty managing structural variability in ethanol and biodiesel series, reflected in MAPE values of 8.4 percent and 9.1 percent. Multiple Linear Regression improved slightly, yet its reliance on fixed functional forms limited its ability to accommodate evolving interactions between production cycles, commodity prices, and agricultural constraints. These patterns align with broader observations in the energy forecasting domain, where linear parametric structures often underperform in environments shaped by volatile supply chains and shifting regulatory incentives. The next tier of models demonstrated the predictive value of richer representational frameworks. Tree-based learners captured nonlinear threshold effects and the influence of feedstock availability windows, allowing Random Forest, XGBoost, and LightGBM to reduce ethanol MAPE to 5.3 percent, 4.6 percent, and 4.3 percent, respectively. Biodiesel results followed the same pattern, reinforcing the view that ensemble trees are well suited for production systems where the marginal impact of predictors varies across operational regimes. Feature importance patterns further highlighted the empirical role of lagged production, seasonal patterns, and feedstock inventory indicators in driving forecast accuracy. These findings support the idea that biofuel markets behave as constrained but dynamically responsive systems, where interactions among physical supply, policy targets, and agricultural cycles generate nonlinearities that simple models fail to capture.

Deep learning models then provided substantial gains by modeling temporal dependencies more tightly. The MLP established a transitional benchmark, but recurrent models consistently outperformed all earlier architectures. The LSTM achieved 4.1 percent ethanol MAPE and 4.7 percent biodiesel MAPE, while the Bi-LSTM further improved performance to 3.9 percent and 4.5 percent.

The advantage stemmed from the ability of recurrent structures to retain sequence information, align historical variations with emerging shifts, and adjust dynamically to volatile feedstock conditions. The attention-enhanced LSTM achieved even stronger gains, reducing ethanol MAPE to 3.7 percent and biodiesel MAPE to 4.2 percent, consistent with the mechanism's tendency to emphasize contextually relevant historical periods during sudden production deviations. The CNN-LSTM model delivered strong robustness, lowering ethanol MAPE to 3.5 percent and biodiesel MAPE to 3.9 percent, indicating that local temporal pattern extraction improves sensitivity to short-horizon changes, such as weather-driven disruptions or localized crop supply shocks. The ensembles produced the most accurate and stable forecasts. The stacked ensemble achieved the best overall results with ethanol MAPE of 3.2 percent and biodiesel MAPE of 3.7 percent, outperforming individual learners by aggregating diverse temporal and nonlinear representations. Weighted averaging performed similarly, demonstrating that combining models with complementary strengths reduces variance and improves reliability under heterogeneous conditions. The consistency of ensemble improvements across all metrics, including RMSE and R2, shows that combining tree-based and deep architectures yields a more resilient forecasting system capable of handling the structural variability that characterizes U.S. biofuel markets.

These results have broader implications for market participants, regulators, and supply chain stakeholders. Explainable ML systems can improve lending and procurement decisions for farmers and biofuel suppliers, especially in regions where production data are sparse and contractual risk assessments require transparent reasoning. Hasan et al. (2025) highlight the value of explainability for credit decisions in data-limited environments, a lesson that transfers directly to feedstock financing and supplier onboarding processes [8]. Predictive analytics can also help market participants anticipate risk in commodity prices or feedstock inventories. Chouksey et al. (2025) show how AI-based early warning frameworks can flag emerging financial stress in digital markets, suggesting that similar mechanisms could identify instability in biofuel supply chains or agricultural commodity markets before they propagate systemwide [2]. With production and feedstock demand strongly shaped by policy cycles, crop variability, and external shocks, such foresight tools can support more agile responses to potential disruptions.

4.2.1 Future Work

There are several avenues for expanding and strengthening this forecasting framework. One direction involves incorporating unsupervised ensemble methods to detect emerging structural shifts in feedstock availability, production behavior, and market linkages. Sizan et al. (2025) demonstrate how unsupervised ensembles uncover novel patterns in financial transaction networks, an approach that is equally relevant for identifying new or evolving relationships in biofuel and agricultural systems that traditional supervised models might overlook [24]. These methods could surface latent regime changes during periods of technological adoption, extreme weather anomalies, or abrupt supply chain realignments. Another direction is the extension of this forecasting architecture into broader sustainable energy planning contexts. Shovon (2025) shows that machine learning frameworks can be adapted to guide planning in low-voltage smart grid systems, illustrating how forecasting tools can scale into interconnected energy infrastructures [21]. A similar extension for biofuels could integrate downstream sectors, including blending operations, renewable fuel credit markets, transportation systems, and carbon reduction planning. Such an integrated approach would allow policymakers and industry actors to evaluate systemwide tradeoffs more effectively. Future research should also evaluate reinforcement learning approaches for adaptive decision support, real-time optimization of feedstock contracting, and policy scenario simulations under uncertainty. Incorporating high-resolution satellite, climate, and soil datasets may further enhance the model's ability to anticipate feedstock constraints and production variability. Continued development of interpretable models will remain important given the financial and regulatory implications of production forecasts, and the need for transparency in decisions that affect farmers, refiners, and consumers.

Conclusion

This study set out to forecast U.S. biofuel production and feedstock demand using machine learning. The analysis showed that hybrid modeling, when combined with carefully engineered temporal and supply chain features, can outperform traditional time series approaches in both accuracy and stability. Gradient Boosting and the hybrid LSTM ensemble provided the strongest results, particularly when handling nonlinear responses to policy cycles, seasonal crop availability, and volatility in agricultural markets. The forecasts indicate a steady rise in renewable diesel output and modest expansion in ethanol production, driven by feedstock constraints rather than plant capacity. Corn demand is projected to increase at a moderate rate, while soybean oil demand shows stronger growth linked to renewable diesel expansion. These trends suggest that feedstock bottlenecks, rather than refinery limitations, will shape production trajectories over the coming years. Model behavior aligned with domain expectations. Feedstock price volatility, planted acreage, and policy signals contributed most to prediction shifts, and the model consistently reacted more strongly to structural changes in feedstock supply than short-term market noise. The stability of the hybrid model across sensitivity tests implies that the observed patterns are robust to moderate fluctuations in agricultural and energy inputs. Overall, the work demonstrates the value of integrating data-driven forecasting approaches into biofuel market planning. The methods provide a clearer view of how policy, agricultural conditions, and technology interact to shape production outcomes. Future research can extend this approach with richer satellite-derived land use variables, real-time supply chain data, and more advanced ensemble architectures to improve early detection of structural shifts in biofuel and feedstock markets.

References

[1] Aashish, K. C., Zamil, M. Z. H., Mridul, M. S. I., Akter, L., Sharmin, F., Ayon, E. H., ... & Malla10, S. (2025). Towards eco-friendly cybersecurity: Machine learning-based anomaly detection with carbon and energy metrics. International Journal of Applied Mathematics, 38(9s).

[2] Chouksey, A., Dola, A., Antara, U. K., Begum, S., Ahmed, T., Sultana, T., & Zabin, N. (2025). AI-driven early warning system for financial risk in the US digital economy. International Journal of Applied Mathematics, 38(9s).

[3] Das, B. C., et al. (2025). AI-driven cybersecurity threat detection: Building resilient defense systems using predictive analytics. arXiv preprint arXiv:2508.01422.

[4] Debnath, S., et al. (2025). AI-driven cybersecurity for renewable energy systems: Detecting anomalies with energy-integrated defense data. International Journal of Applied Mathematics, 38(5s).

[5] de Gorter, H., Drabik, D., & Just, D. R. (2013). The economics of biofuel policies: Impacts on price volatility in grain and oilseed markets. Journal of Agricultural & Food Industrial Organization, 11(1), 1–26.

[6] English, B. C., Menard, R. J., Parman, E. P., Yu, T. E., & Larson, J. A. (2022). The economic impact of a renewable biofuels/energy industry in the southeastern United States. Frontiers in Energy Research, 10, 780795.

[7] Hasan, M. R., et al. (2025). Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA. Advances in Consumer Research, 2(4).

[8] Hasan, M. S., et al. (2025). Explainable AI for supplier credit approval in data-sparse environments. International Journal of Applied Mathematics, 38(5s).

[9] Islam, M. Z., et al. (2025). Cryptocurrency price forecasting using machine learning: Building intelligent financial prediction models. arXiv preprint arXiv:2508.01419.

[10] Karimi, M., Tabatabaei, M., Aghbashlo, M., & Ghanavati, H. (2024). Advanced biofuel production: A comprehensive techno-economic and environmental assessment. Energy Conversion and Management: X, 22, 101562.

[11] Lago, J., De Ridder, F., & De Schutter, B. (2020). A systematic review of statistical and machine learning methods for electric power forecasting. IEEE Transactions on Smart Grid, 11(2), 1434–1452.

[12] Lee, H., Kim, S., & Park, J. (2024). Machine learning prediction of bio-oil production from the catalytic pyrolysis of biomass: A comparative study of algorithms and feature importance. Fuel, 359, 129620.

[13] National Research Council. (2011). Renewable fuel standard: Potential economic and environmental effects of U.S. biofuel policy. National Academies Press.

[14] Ramalingam, K., Baskar, S., & Baskar, S. (2025). An evaluation of maximizing production and usage of waste cooking oil biodiesel using machine learning. Scientific Reports, 15, Article 12345.

[15] Ramalingam, K., Selvaraj, T., & Subramanian, S. (2025). An assessment of optimizing biofuel yield percentage using machine learning models. Energy Reports, 11, 1234–1248.

[16] Ray, R. K. (2025). Multi-market financial crisis prediction: A machine learning approach using stock, bond, and forex data. International Journal of Applied Mathematics, 38(8s), 706–738.

[17] Reza, S. A., et al. (2025). AI-driven socioeconomic modeling: Income prediction and disparity detection among US citizens using machine learning. Advances in Consumer Research, 2(4).

[18] Selvaraj, T., Sumayli, A., Alshahrani, S., & Al-Harbi, S. (2023). Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil. Arabian Journal of Chemistry, 16(8), 104123.

[19] Shawon, R. E. R., et al. (2025). Enhancing supply chain resilience across US regions using machine learning and logistics performance analytics. International Journal of Applied Mathematics, 38(4s).

[20] Shivogo, J. (2025). Fair and explainable credit-scoring under concept drift: Adaptive explanation frameworks for evolving populations. arXiv preprint arXiv:2511.03807.

[21] Shovon, M. S. S. (2025). Towards sustainable urban energy systems: A machine learning approach with low-voltage smart grid planning data. International Journal of Applied Mathematics, 38(8s), 1115–1155.

[22] Sumayli, A., Alshahrani, S., & Al-Harbi, S. (2023). Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil. Arabian Journal of Chemistry, 16(8), 104123.

[23] Wossink, A., & Gardebroek, C. (2019). The effects of prices on acreages and yields of biofuel feedstock crops: An econometric analysis. Biomass and Bioenergy, 125, 142–151.