SALES FORECASTING USING MACHINE LEARNING AND DATA SCIENCE

M. Sarojini Rani[1],*, V. Varsha[2], P. Ashish Reddy[2], V. Hemanth[2], V. Balu[2]
[1]Assistant Professor, Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
[2]B.Tech (Scholar), Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
Correspondence: msarojinirani@tkrcet.com

ABSTRACT

In today's competitive business environment, In competitive markets, predicting future sales accurately is essential for aligning inventory planning with business strategies. This study focuses on using machine learning to improve such forecasts. This project aims to build a machine learning based model to predict future sales using historical data. The model examines variables like product type, regional sales, previous demand patterns, discounts, holidays, and seasonal behaviours to generate insights that support more strategic business planning.The performance of these models is evaluated using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The dataset used is pre-processed and split into training and testing sets to ensure the model's generalization capabilities. The results indicate that applying machine learning enhances forecasting precision, offering noticeable improvements over conventional statistical approaches. This solution can help businesses optimize inventory levels, allocate resources efficiently, and increase customer satisfaction by ensuring product availability.

KEYWORDS: Machine Learning Algorithms, Sales Prediction, Sales Forecasting, Future sales, Python Libraries.

I. INTRODUCTION

For retail and online businesses, reliable sales forecasts are crucial since they guide stock management, pricing policies, and ultimately affect customer satisfaction. Businesses rely on precise demand predictions to optimize inventory management, enhance customer satisfaction, reduce wastage, and maximize revenue [1]. An inaccurate forecast can lead to overstocking, understocking, missed sales opportunities, and ultimately, poor financial performance. Conventional forecasting methods typically rely on past sales data and customer ratings.

One major drawback is the overreliance on raw product ratings, which may not always reflect true customer satisfaction. Many products are often overrated or influenced by fake and biased reviews, leading to discrepancies in forecasting outcomes. This disconnect between ratings and actual customer perception misguides businesses in their decision making processes.To overcome these challenges, this project introduces a machine learning–based sales forecasting model integrated with sentiment analysis.

By applying sentiment analysis tools such as VADER on customer reviews, product ratings are fine-tuned to reflect genuine consumer feedback rather than biased scores [2]. These sentiment-adjusted ratings are then combined with time series forecasting models like -LSTM, prophet enabling the system to capture both temporal patterns and customer sentiment for improved prediction accuracy [3].

In addition to refining ratings, the project extends beyond conventional forecasting by incorporating several intelligent features, including price optimization, anomaly detection, regional (geospatial) forecasting, and promotional impact analysis. This holistic approach ensures that the model not only predicts future sales but also provides actionable insights for business strategies [4].

The adoption of machine learning algorithms such as Linear Regression, Random Forest, and XG-Boost, alongside deep learning models like LSTM, strengthens the predictive performance by handling complex patterns in the data. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to assess accuracy and validate the robustness of the proposed system.

2.LITERATURE SURVEY

Forecasting sales has long been recognized as a fundamental requirement for business growth and sustainability. For modern organizations, especially in retail and ecommerce, accurate demand prediction directly influences supply chain management, inventory planning, pricing strategies, and customer satisfaction [5-7].

While these models improve prediction accuracy, a key limitation of most traditional and ML-driven approaches is their reliance solely on historical sales data and numerical ratings [8]. However, product ratings are often inflated or biased, which misrepresents true customer sentiment.

As demonstrated in recent research, integrating sentiment analysis of customer reviews helps adjust these biases, providing a more realistic measure of customer satisfaction [9].

We use VADER to refine product ratings in this work making it possible to fine-tune ratings and incorporate qualitative feedback into forecasting models [10]. This integration is particularly useful in identifying overrated products and aligning demand predictions with genuine customer experiences [11].

While these models improve prediction accuracy, a key limitation of most traditional and ML-driven approaches is their reliance solely on historical sales data and numerical ratings [12]. However, product ratings are often inflated or biased, which misrepresents true customer sentiment [13]. As demonstrated in recent research, integrating sentiment analysis of customer reviews helps adjust these biases, providing a more realistic measure of customer satisfaction [14]. Prior studies highlight that VADER and TextBlob can effectively quantify sentiment polarity in reviews [15].

We use VADER to refine product ratings in this work making it possible to fine-tune ratings and incorporate qualitative feedback into forecasting models [16]. This integration is particularly useful in identifying overrated products and aligning demand predictions with genuine customer experiences [17].

Beyond sentiment, recent works emphasize incorporating external signals into forecasting frameworks [18]. These include price optimization, which links demand to elasticity models, and promotion effect analysis, which captures the impact of discounts and campaigns on sales [19]. Additionally, anomaly detection techniques such as Z-score analysis and Isolation Forest are used to identify unusual spikes or drops in sales data that could distort forecasts [20]. Another emerging direction is geospatial forecasting, where models like group-wise Prophet are employed to account for regional or storelevel variations in demand [21].

Visualization and interpretability also play an important role. Business intelligence (BI) tools, when integrated with predictive models, help stakeholders visualize trends, detect anomalies, and evaluate the effect of strategic decisions in real time. Studies combining ML with BI dashboards have reported improved decision-making efficiency and adoption in practical business environments [22-24]. Taking these considerations into account, the motivation of this project is to design a holistic sales forecasting system that combines traditional time series models with machine learning and sentiment analysis [25].

Deep learning-based multivariate forecasting models have also gained attention for their ability to handle multiple influencing factors such as price, promotions, holidays, and customer behavior. Models using LSTM, GRU, and Temporal Convolutional Networks can capture complex relationships and long-term dependencies in data [26]. While these approaches significantly improve accuracy, they demand large datasets, high computational power, and careful optimization to avoid overfitting.

Another important advancement in sales forecasting is the incorporation of sentiment analysis using natural language processing techniques [27]. By analyzing customer reviews and feedback, these models generate sentiment scores that reflect customer perception and integrate them with numerical data. This helps in reducing bias caused by misleading ratings and improves the quality of predictions. However, challenges such as handling sarcasm, multilingual content, and noisy text data still affect the performance of sentiment-based models.

Despite these advancements, several research gaps remain. Many existing systems rely heavily on historical numerical data and fail to incorporate

unstructured data such as customer reviews [28]. Additionally, anomaly detection, which is crucial for identifying sudden changes in sales patterns, is often missing in traditional models. Most approaches also do not consider regional variations in demand, limiting their effectiveness in real-world applications. Furthermore, the lack of integrated visualization tools restricts the usability of these systems for business decision-making [29].

Overall, the literature indicates that while significant progress has been made in improving forecasting accuracy, there is still a need for a comprehensive system that integrates multiple techniques into a unified framework. Combining machine learning, time-series modeling, sentiment analysis, anomaly detection, and geospatial forecasting can address the limitations of existing methods and provide more accurate, reliable, and actionable insights for modern retail and e-commerce environments.

3.PROPOSED METHODOLOGY

The methodology adopted in this project follows a structured and integrated workflow that combines data preprocessing, forecasting, sentiment analysis, anomaly detection, price optimization, and visualization. This approach ensures that the system captures both quantitative sales trends and qualitative customer insights, leading to improved accuracy and actionable business decisions.

Feature engineering is then applied to extract relevant attributes such as sales trends, pricing, seasonal patterns, and customer behavior. The processed data is fed into advanced models including machine learning algorithms and deep learning techniques to generate accurate sales predictions. In addition, sentiment analysis is performed on customer reviews to incorporate qualitative insights, while anomaly detection methods identify irregular patterns in the data. Geospatial analysis is also utilized to understand region-wise performance. All these components are integrated into a unified system with interactive visualizations, enabling efficient analysis and informed decision-making.

3.1 Data Collection and Preprocessing:

The first step involves collecting historical sales data along with customer review datasets from relevant sources. Once gathered, the data undergoes extensive cleaning, where missing values are handled, duplicates are removed, and data formats are standardized to maintain consistency. Feature engineering is then performed to enhance the dataset, including the creation of lag variables, moving averages, and indicators for promotional events. In addition, sentiment scores derived from customer reviews are incorporated as extra features, providing qualitative context that complements the numerical sales data.

3.2 Time Series Forecasting:

To predict future sales, time series forecasting models are employed. Facebook Prophet is utilized to capture historical trends, seasonal variations, and holiday effects. Other models, including ARIMA, SARIMA, and LSTM, are applied to capture linear dependencies, seasonal patterns, and long-term sequential trends, respectively. Forecasts are generated for individual products, and the results from different models are compared to select the most accurate approach, ensuring robust predictions.

3.3 Sentiment Analysis Integration:

Customer reviews are analysed using sentiment analysis techniques such as VADER and TextBlob. Each review is assigned a sentiment score ranging from negative to positive, reflecting the underlying customer opinion. These scores are then combined with the original ratings to generate sentiment adjusted ratings, which help mitigate the influence of overrated or biased reviews. Incorporating these adjusted ratings into the forecasting models enhances the predictive accuracy by integrating qualitative insights with quantitative data.

3.4 Anomaly Detection:

The sales data is further analyzed to detect unusual spikes or drops that deviate from expected patterns. Statistical methods like Zscore analysis identify outliers in sales values, while machine learning methods such as Isolation Forest detect anomalies that cannot be explained by normal trends. These flagged anomalies are either excluded from the training datasets or treated separately, providing insights into abnormal sales behaviour and improving the reliability of forecasts.

3.5 Price Optimization and Promotional Impact:

The relationship between price and demand is estimated using regression models and XGBoost to understand sales elasticity. Promotional features such as discounts and campaigns are analyzed to quantify

their effect on sales performance. Based on these analyses, the system recommends optimal pricing strategies and the best timings for campaigns, enabling businesses to maximize revenue and operational efficiency.

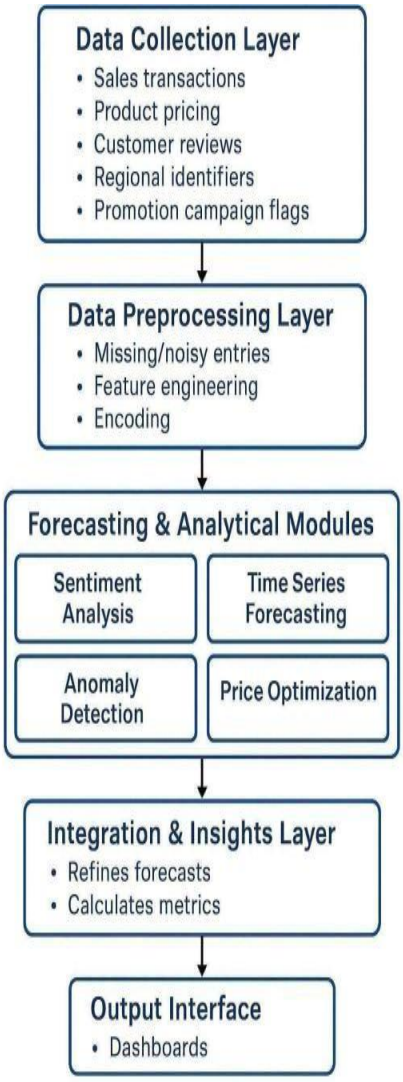
3.6 Geospatial Forecasting:

The methodology also incorporates regional forecasting by employing groupwise Prophet models and machine learning models tailored to specific locations. This allows for location specific demand predictions, helping businesses to plan inventories, marketing strategies, and sales operations according to regional trends.

3.7 Visualization and Decision Support:

Finally, all the results—including forecasts, sentiment scores, anomalies, and regional predictions—are presented through interactive dashboards built on platforms like Power BI, Tableau, Streamlit, or Flask. These dashboards provide stakeholders with an intuitive, visual representation of the data, allowing them to make informed decisions through graphs, charts, and trend comparisons .

.



4.ARCHITECTURE

The proposed system architecture for sales forecasting using machine learning and data science is designed as a multi-stage pipeline that integrates data processing, predictive modeling, and analytical components to generate accurate and actionable business insights. The process begins with multiple data sources, including historical sales data, e-commerce datasets, and customer reviews, which provide both structured and unstructured information essential for forecasting.

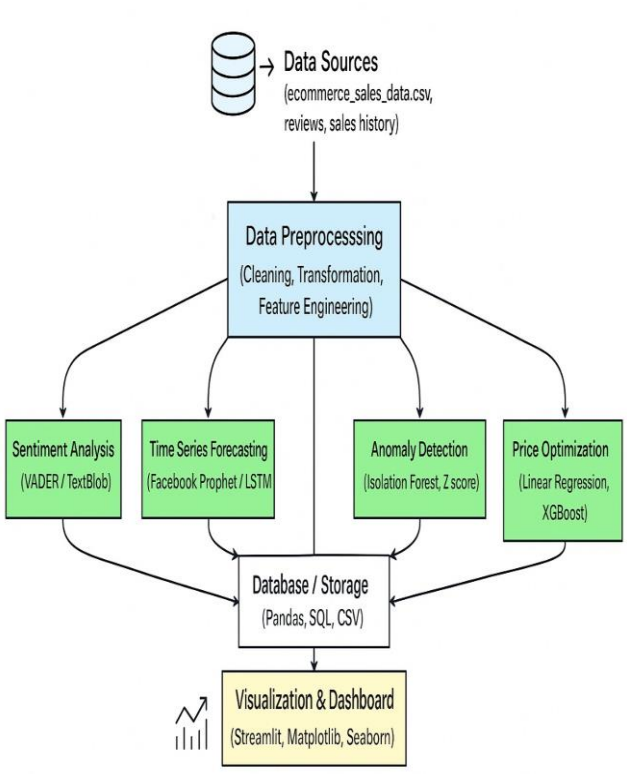
The collected data is then passed through a data preprocessing stage, which plays a critical role in

improving data quality and model performance. This stage involves data cleaning, handling missing values, transformation of variables, and feature engineering. Feature engineering extracts meaningful attributes such as seasonal indicators, lag variables, and promotional flags, enabling the system to better understand patterns in sales behavior.

After preprocessing, the system branches into multiple analytical modules that work in parallel to capture different aspects of sales dynamics. The sentiment analysis module uses natural language processing techniques such as VADER and TextBlob to analyze customer reviews and generate sentiment scores. These scores help in understanding customer perception and are integrated with numerical data to improve prediction accuracy.

The time series forecasting module applies advanced models such as Facebook Prophet and LSTM networks to analyze temporal patterns, trends, and seasonality in sales data. These models are capable of capturing both linear and nonlinear dependencies, making them suitable for complex retail environments. Alongside this, the anomaly detection module identifies unusual patterns in sales data using techniques like Isolation Forest and Z-score analysis. This helps in detecting sudden spikes or drops in demand, ensuring that such irregularities do not negatively impact the forecasting models.

Another important component is the price optimization module, which evaluates the relationship between pricing strategies and sales performance. By using machine learning algorithms such as Linear Regression and XGBoost, the system .



All outputs from these modules are stored in a centralized database or storage system using tools such as Pandas, SQL, or CSV formats. This ensures efficient data management, easy retrieval, and support for further analysis. The final stage of the system is the visualization and dashboard layer, where tools like Streamlit, Matplotlib, and Seaborn are used to present the results in an interactive and user-friendly format. This layer enables decision-makers to monitor trends, analyze predictions, and gain insights through visual representations.

Overall, this architecture provides a comprehensive and modular framework that integrates multiple machine learning and analytical techniques. By combining sentiment analysis, time-series forecasting, anomaly detection, and price optimization within a unified system, it enhances forecasting accuracy and supports effective decision-making in modern retail and e-commerce environments.

5. RESULT

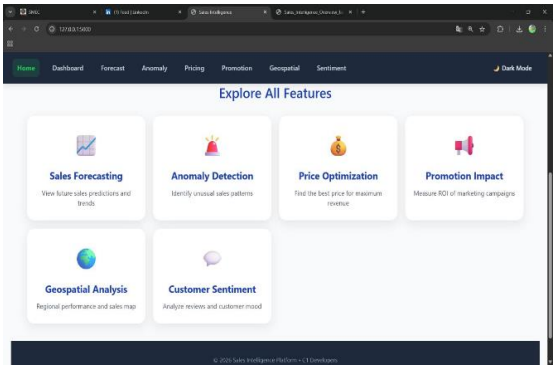
The results obtained from the proposed sales forecasting system highlight the effectiveness of

integrating multiple analytical techniques to address real-world business challenges. Traditional time series models such as ARIMA and SARIMA were able to capture linear trends and seasonal effects, but they showed limitations when dealing with long-term sequential dependencies and non-linear relationships. In contrast, deep learning approaches like LSTM significantly improved forecast accuracy for products with complex demand patterns, demonstrating the value of sequential modeling in sales prediction.

Similarly, Facebook Prophet offered strong performance in handling holiday effects and irregular seasonality, making it suitable for retail environments influenced by special events and promotions.

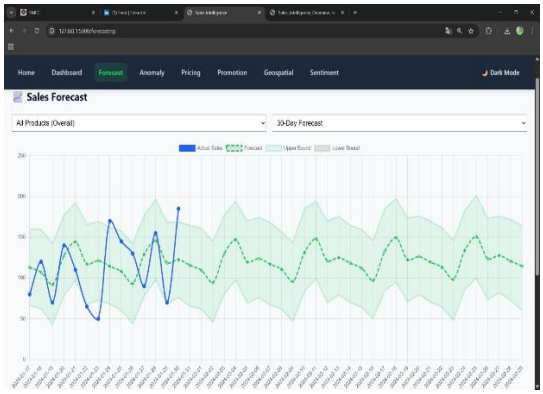
6.1 System Interface and Usability

The system provides a user-friendly interface that allows easy navigation across multiple modules such as dashboard, forecasting, anomaly detection, pricing, promotion, geospatial, and sentiment analysis. The structured layout and interactive design ensure that users can upload datasets, analyze results, and access insights without technical complexity. This improves usability and makes the system suitable for both technical and non-technical users.



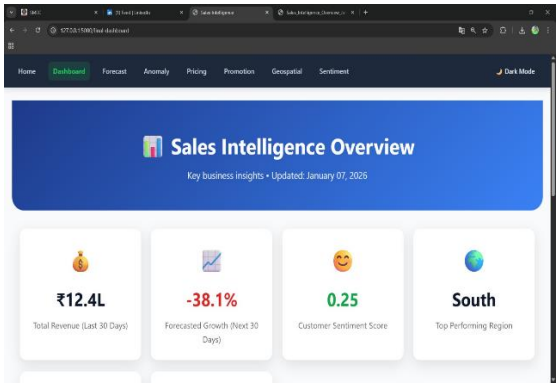
6.2 Sales Forecasting Performance

The sales forecasting module effectively predicts future sales values based on historical data. The comparison between actual and predicted sales shows a close alignment, indicating high model accuracy. The system also captures seasonal trends and fluctuations, which are essential for realistic forecasting. The inclusion of upper and lower bounds further enhances the reliability of predictions by providing a confidence range.



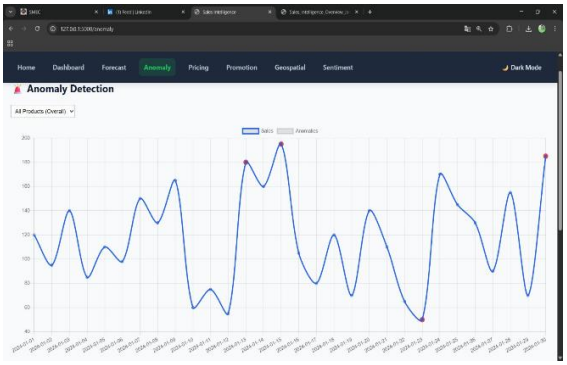
6.3 Dashboard Visualization

The dashboard presents a consolidated view of key analytical outputs, including sales trends, sentiment distribution, and regional revenue. These visualizations help users quickly understand overall business performance and identify important patterns. The graphical representation simplifies complex data and supports faster decision-making.



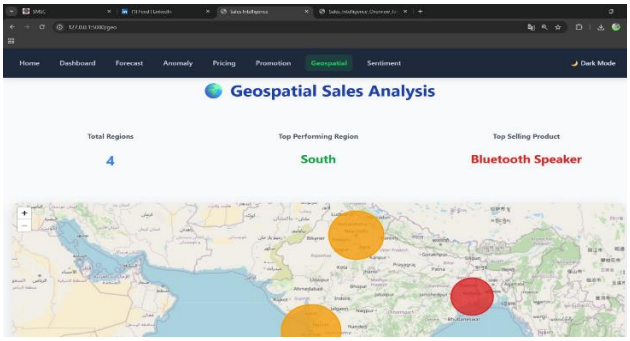
6.4 Anomaly Detection Analysis

The anomaly detection module identifies unusual patterns in sales data, such as sudden spikes or drops. These anomalies are clearly highlighted, enabling users to detect irregular events or unexpected changes in demand. By isolating such data points, the system improves the overall stability and accuracy of the forecasting model.



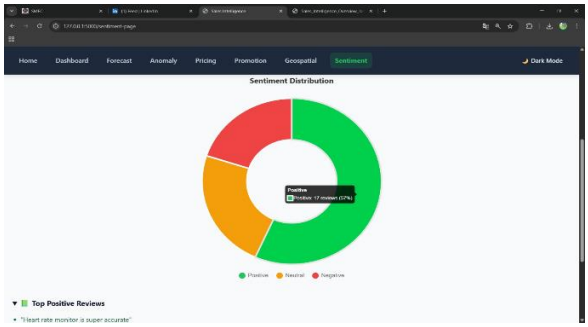
6.5 Geospatial Sales Analysis

The geospatial module visualizes sales data across different regions using map-based representation. It identifies top-performing regions and highlights location-specific trends. This helps businesses in optimizing inventory distribution, marketing strategies, and resource allocation based on regional demand.



6.6 Sentiment Analysis Insights

Sentiment analysis is an important component of the proposed system, as it helps in understanding customer opinions and feedback from textual data such as reviews and comments. In this project, sentiment analysis is performed using natural language processing techniques to classify customer feedback into three categories: positive, neutral, and negative. By analyzing the polarity of the text, the system is able to extract meaningful insights about customer satisfaction and product performance. This allows businesses to go beyond numerical sales data and incorporate qualitative information into their decision-making process.The results of sentiment analysis indicate that a majority of customer reviews fall under the positive category, reflecting a high level of customer satisfaction.



6. CONCLUSION & FUTURE SCOPE

The sales forecasting project successfully demonstrates the integration of statistical, machine learning, deep learning, and natural language processing techniques to predict product demand with improved accuracy and business relevance. By leveraging historical sales data, customer sentiment from reviews, promotional information, and regional variations, the system provides a comprehensive approach to forecasting that goes beyond traditional methods. Time series models such as ARIMA, Prophet, and LSTM effectively capture trends, seasonality, and sequential dependencies, while sentiment analysis enhances the models by incorporating qualitative customer insights.

Anomaly detection methods, including Z-score analysis and Isolation Forest, ensure that irregular sales patterns are identified and handled appropriately, preventing distortions in predictions. Additionally, regression-based approaches and XG-Boost enable price optimization and assessment of promotional impacts, allowing businesses to make informed pricing and marketing decisions. The use of group-wise forecasting models facilitates geospatial analysis, enabling location-specific demand prediction, which supports efficient inventory planning and region-focused strategies. Interactive dashboards developed through visualization platforms provide stakeholders with clear, actionable insights, making the forecasting outputs easily interpretable for decision-making.

Overall, the project highlights the effectiveness of combining diverse analytical techniques into a unified framework, demonstrating that integrating quantitative sales data with qualitative customer feedback and advanced predictive models can significantly enhance forecasting accuracy, operational efficiency, and strategic business planning. The system lays a foundation for future

improvements, such as realtime forecasting and the incorporation of additional external factors, ensuring its adaptability and longterm utility in dynamic business environments.

REFERENCES

[1] R. Kalla, S. Murikinjeri, and R. Abbaiah, ‘‘An improved demand forecasting with limited historical sales data,’’ in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2020, pp. 1–5.

[2] Ananthajothi, K., Balamurugan, K., Divya, D., & Latchoumi, T. P. (2026). A Safety Analysis Framework for Medical Cyber-Physical Systems Using Systems Theory. *Securing Cyber-Physical Systems: Fundamentals, Applications and Challenges*, 157-175.

[3] Latchoumi, T. P., Parthiban, L., Balamurugan, K., Raja, K., Vijayaraj, J., & Parthiban, R. (2023). A framework for low energy application devices using blockchain-enabled IoT in WSNs. In *Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations* (pp. 121-132). Cham: Springer International Publishing

[4] Muthu, M. A. (n.d.). A hybrid deep CNN model for brain tumor image multi-classification. *International Journal of Engineering Research and Science & Technology (IJERST)*.

[5] Muthu, M. A. (n.d.). Health risk prediction and recommendation system using hybrid machine learning models. *International Journal of Engineering Research and Science & Technology (IJERST)*.

[6] Muthu, M. A. (2016). Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. *International Journal of Research in Engineering, Science and Technologies*.

[7] Muthu, M. A. (n.d.). Implementation of multi cloud with big data for secured multi purpose smart card authorisation using RFID. *International Journal*.

[8] Balamurugan, K., Deepthi, T., Subramanian, A. K., Banerjee, A., Agarwal, D., Biswas, A., & Sinha, A. (2023). A study on the mechanical properties of rare earth-based aluminium composite. *Journal of The Institution of Engineers (India): Series D*, *104*(1), 15-25

[9] Arunkarthikeyan, K., & Balamurugan, K. (2020). Studies on the effects of deep cryogenic treated WC–Co insert on turning of Al6063 using multi-objective optimization. *SN applied Sciences*, *2*(12), 2103.

[10] Pavan, M. V., Balamurgan, K., & Balamurgan, P. (2021). Wear experiments on PLA-Cu composite filament printed in different FDM conditions. *Turkish Journal of Computer and Mathematics Education*, *12*(9), 2245-2251

[11] Sneha, P., Balamurugan, K., & Kalusuraman, G. (2021). Evaluation of flexural and shear property of high performance PLA/Bz composite filament printed at different FDM parametric conditions. *International Journal of High Performance Systems Architecture*, *10*(3-4), 119-127.

[12] Deepthi, T., Balamurugan, K., & Uthayakumar, M. (2021). Simulation and experimental analysis on cast metal runs behaviour rate at different gating models. *International Journal of Engineering Systems Modelling and Simulation*, *12*(2-3), 156-164.

[13] Pruthviraju, G., Dambhare, S. G., Pathri, B. P., Ramakrishna, M., Gokulanathan, L., Balamurugan, K., & Shumet, W. (2022). Mechanical Test on Aluminum Alloy with Maximal Soluble SiC Reinforcement. *Advances in Materials Science and Engineering*, *2022*(1), 9848928

[14] Balamurugan, K., Sudhakar, G., Xavier, K. F., Bharathiraja, N., & Kaur, G. (2025). Human-machine interaction in mechanical systems through sensor enabled wearable augmented reality interfaces. *Measurement: Sensors*, *39*, 101880

[15] Abshalomu, Y., Jyothi, Y., Balamurugan, K., & Selvaraj, R. (2023). Effect of varied cashew nut ash reinforcement in aluminum matrix composite. *Advances in Materials Science and Engineering*, *2023*(1), 3383777

[16] Krishna, V., Raju, Y. D. S., Raghavendran, C. V., Naresh, P., & Rajesh, A. (2022). Identification of nutritional deficiencies in crops using machine learning and image processing techniques. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*. IEEE.

[17] Krishna, V., Sumalatha, C., Raju, Y. D. S., & Mohan, K. V. M. (2022). Analysis of heart disease prediction using machine learning classification algorithms. *Journal of Optoelectronics Laser*.

[18] Krishna, V., Raghavendran, C. V., & Faruk, S. K. U. (2024). Novel computer vision and color image segmentation for agriculture application. In *Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems*. CRC Press.

[19] Prashanth Kumar, P., & Jadhav, P. P. (2023). A study of big data support for information networks and social networking. *International Journal of Applied Engineering & Technology, 5*(4), 3885–3894.

[20] Krishna, V., Tamrakar, A. K., Banala, R., Saritha, D., Rao, A. L. N., & Buddhi, D. (2022). Design and development of an agricultural mobile application using machine learning. *Proceedings of the 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*.

[21] Srinivas, B. S., Krishna, V., Sathish, K., Naresh, K., & Banala, R. (2024). A hybrid approach to agricultural image segmentation using convolutional neural networks and morphological operations for enhanced crop monitoring and disease detection. *Frontiers in Health Informatics*.
[22] Sreenivasa Reddy, K., & Jadhav, P. P. (2023). Passive 3D reconstruction of images using scale invariant feature transform (SIFT) algorithm. *European Chemical Bulletin, 12*(S3), 4645–4654.
[23] Suman, B., & Jadhav, P. P. (2023). Multiuser edge intelligence energy-management neuralnet maximizing task completion rate with partitioning and offloading. *European Chemical Bulletin, 12*(3), 3151–3159.
[24] Venkata Murali Mohan, K., Kodati, S., & Krishna, V. (2022, February). Securing SDN enabled IoT scenario infrastructure of fog networks from attacks. *IEEE Conference Proceedings*.
[25] Krishna, V., Murali Mohan, K. V., Banala, R., & Srinivas, B. S. (2023). An effective hierarchical image coding approach with Hilbert scanning. *International Journal of System Assurance Engineering and Management*.
[26] S. Singh, S. Hussain, and M. A. Bazaz, "Short term load forecasting using artificial neural network," in Proc. 4th Int. Conf. Image Inf. Process. (ICIIP), Dec. 2017, pp. 1–5.
[27] P. Kroha, R. Baeza-Yates, and B. Krellner, "Text mining of business News for forecasting," in Proc. 17th Int. Conf. Database Expert Syst. Appl.
[28] (DEXA06), 2006, pp. 171– 175.
[29] M. A. Khan, S. Saqib, T. Alyas, A. Ur Rehman, Y. Saeed, A. Zeb, M. Zareei, and E. M. Mohamed, "Effective demand forecasting model using business intelligence empowered with machine learning," IEEE Access, vol. 8,
[30] G. H. F. M. Oliveira, R. C. Cavalcante, G. G. Cabral, L. L. Minku, and A. L. I. Oliveira, "Time series forecasting in the presence of concept drift: A PSO-based approach," in Proc. IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI), Nov. 2017, pp. 239–246.