Power System Protection and Control
ISSN:1674-3415

PREDICTIVE CAPACITY PLANNING USING MACHINE LEARNING FOR ENTERPRISE APPLICATIONS

Gunjan Shegade
gunjanshegade@gmail.com

Recevied:29 June 2024 **Revised**: 22 July 2024 **Accepted**: 20 August 2024

ABSTRACT

Capacity planning for enterprise applications has traditionally relied on rule-of-thumb heuristics, static provisioning based on peak estimates, and reactive scaling that responds to problems only after they manifest as performance degradation. These approaches lead to either costly over-provisioning or risky under-provisioning, and neither adapts well to the increasingly dynamic and unpredictable workloads of modern enterprise systems. This paper presents a machine learning framework for predictive capacity planning that forecasts resource demand ahead of time, enabling proactive provisioning that balances cost and performance. We designed a forecasting system combining time-series models for baseline demand prediction with gradient boosted models for incorporating business drivers, and evaluated it against conventional capacity planning approaches using 18 months of production telemetry from a portfolio of enterprise applications. Results show that the ML framework reduced resource over-provisioning by 32% while simultaneously reducing capacity-related performance incidents by 58% compared to static peak-based provisioning. Forecast accuracy for CPU and memory demand achieved mean absolute percentage errors of 11.4% and 9.7% respectively at a two-week horizon. The framework's incorporation of business event calendars improved forecast accuracy during high-variance periods such as month-end processing and marketing campaigns by 27%. The paper discusses the practical integration of predictive capacity planning into enterprise operations, including the organizational shift from reactive to proactive capacity management, and addresses the challenges of forecast uncertainty, cold-start for new applications, and model maintenance.

keywords: *Capacity Planning, Machine Learning, Demand Forecasting, Resource Provisioning, Enterprise Applications, Time-Series Forecasting, Proactive Scaling*

INTRODUCTION

Every enterprise application consumes computing resources, and someone must decide how much of those resources to provision. This decision, capacity planning, has enormous financial and operational consequences [1]. Provision too much, and the organization pays for idle infrastructure that delivers no value. Provision too little, and applications slow down, fail, or become unavailable during demand peaks, damaging user experience and business operations. Getting capacity planning right is a persistent challenge that grows harder as application portfolios expand and workloads become more variable [2].

Traditional capacity planning approaches have significant limitations. Static provisioning based on estimated peak demand is simple but wasteful, since resources sized for the annual peak sit largely idle the rest of the time [3]. Reactive auto-scaling, which adds resources when utilization crosses a threshold, responds only after demand has already risen, meaning users experience degraded performance during the lag between demand increase and capacity response. Rule-of-thumb heuristics based on historical growth rates fail to capture the complex, multi-driver nature of real enterprise demand, which responds to business cycles, marketing campaigns, product launches, and external events [4].

The promise of predictive capacity planning is to forecast demand before it materializes, enabling proactive provisioning that has capacity in place when it is needed and releases it when it is not [5]. Machine learning is well-suited to this forecasting problem because enterprise demand exhibits patterns, seasonal cycles, weekly rhythms, growth trends, and event-driven spikes, that models can learn from historical telemetry. By combining these learned patterns with knowledge of upcoming business events, a predictive system can anticipate demand far more accurately than reactive or static approaches.

Power System Protection and Control
ISSN:1674-3415

This paper develops and evaluates a machine learning framework for predictive capacity planning, demonstrating that it can simultaneously reduce cost through less over-provisioning and improve reliability through fewer capacity-related incidents, a combination that traditional approaches force into a trade-off.

LITERATURE REVIEW

Capacity planning has a long history in operations research and computer systems performance engineering. Classical approaches drew on queueing theory to model system behaviour under load and to predict the resource levels required to meet performance targets [6]. These analytical models provided valuable insight but required assumptions about workload characteristics that often did not hold in practice, and they struggled to incorporate the business context that drives real demand variation.

The rise of cloud computing transformed capacity planning by making resources elastic and rapidly provisionable. Research on cloud auto-scaling has explored reactive, predictive, and hybrid approaches [7]. Reactive scaling based on utilization thresholds is widely deployed but suffers from the fundamental lag problem. Predictive scaling using time-series forecasting emerged to address this lag, applying methods such as ARIMA, exponential smoothing, and later machine learning to forecast demand ahead of time.

Time-series forecasting methods have been extensively studied for workload prediction. Classical statistical methods including ARIMA and seasonal decomposition capture trend and seasonality well but struggle with irregular, event-driven variation [8]. Machine learning approaches including gradient boosting and neural networks, particularly LSTM and more recently transformer-based models, have shown improved accuracy by capturing complex nonlinear patterns and incorporating exogenous variables. The incorporation of external features such as calendar events and business drivers has been shown to improve forecast accuracy substantially during high-variance periods.

The practical integration of predictive capacity planning into enterprise operations has received less research attention than the forecasting algorithms themselves. Studies have noted that the organizational shift from reactive firefighting to proactive planning requires changes in process and culture beyond the technical capability [9]. Concerns including forecast uncertainty quantification, handling of new applications without historical data, and ongoing model maintenance as workloads evolve have been identified as important for real-world deployment but remain underexplored in the literature.

METHODOLOGY

We designed a two-stage forecasting framework. The first stage uses time-series models to capture baseline demand patterns including trend, seasonality, and autocorrelation. The second stage uses a gradient boosted model to adjust the baseline forecast by incorporating business drivers such as scheduled events, marketing campaigns, and known batch processing windows [14].

The baseline demand forecast at horizon h was modelled as:

$$\hat{y}t + h = Tt + h + S_{t+h} + R_{t+h}$$

where T is the trend component, S is the seasonal component, and R is the residual captured by the time-series model [15].

Forecast accuracy was measured using mean absolute percentage error:

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$Over - provisioning\ was\ quantified\ as\ the\ average\ unused\ capacity\ ratio$:

$$OP = \frac{1}{N}\sum_{i=1}^{N}\frac{C_i - U_i}{C_i} \times 100$$

where C_i is provisioned capacity and U_i is actual utilization at time i [16].

The provisioning decision incorporated forecast uncertainty by provisioning to a confidence bound:

$$C_{provision} = \hat{y}t + h + z\alpha \cdot \sigma_{forecast}$$

where z_\alpha reflects the desired service level and \sigma_{forecast} is the forecast standard deviation [17].

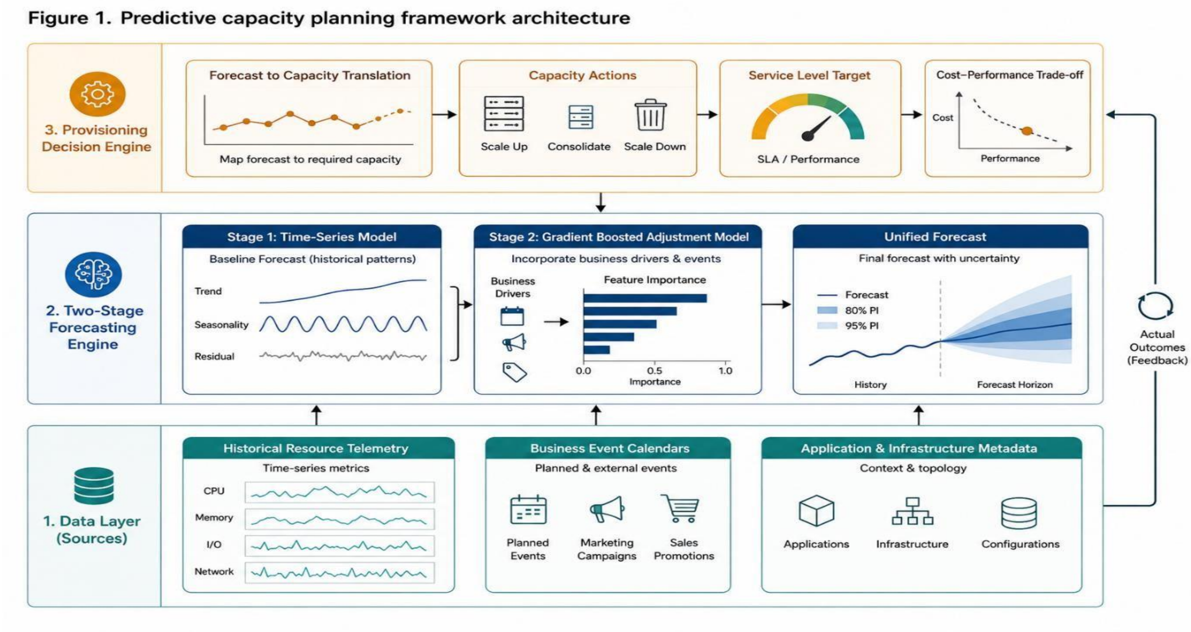

Figure 1. Predictive capacity planning framework architecture.

EXPERIMENTAL SETUP

We evaluated the framework using 18 months of production telemetry from a portfolio of 24 enterprise applications spanning transaction processing, analytics, web services, and batch workloads. Telemetry included CPU utilization, memory usage, I/O rates, and request volumes sampled at five-minute intervals. Business event data included fiscal calendar events, scheduled marketing campaigns, and known batch processing windows [18].

The dataset was split temporally, with the first 12 months used for training and the final 6 months held out for evaluation. Three baseline approaches were compared: static peak-based provisioning sized to the historical maximum, reactive threshold-based auto-scaling, and a simple linear growth projection. Forecast horizons of one day, one week, and two weeks were evaluated to assess accuracy degradation over longer horizons.

Performance was measured across forecast accuracy, over-provisioning ratio, and capacity-related incident count. A capacity-related incident was defined as any period where utilization exceeded 90% of provisioned capacity, risking performance degradation.

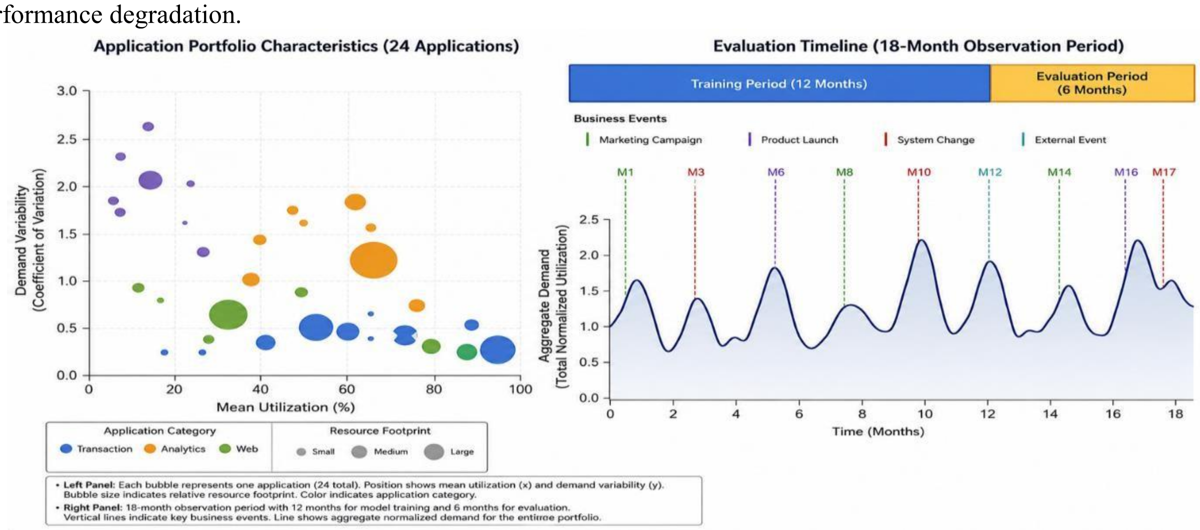

Figure 2. Application portfolio characteristics and evaluation timeline.

Power System Protection and Control
ISSN:1674-3415

Table 1. Enterprise application portfolio characteristics.

Application Category	Count	Mean Utilization	Demand Variability (CoV)	Forecast Difficulty
Transaction processing	8	47%	0.31	Moderate
Analytics/reporting	6	38%	0.62	High
Web services	7	52%	0.28	Low
Batch workloads	3	71%	0.84	Very High
Portfolio total	**24**	**48%**	**0.44**	**Moderate**

RESULTS

5.1 Forecast Accuracy

The ML framework achieved MAPE of 11.4% for CPU demand and 9.7% for memory demand at a two-week horizon. Accuracy was highest for stable web service workloads and lowest for highly variable batch workloads, as expected. Accuracy degraded gracefully with horizon, rising from 6.2% MAPE at one day to 11.4% at two weeks for CPU.

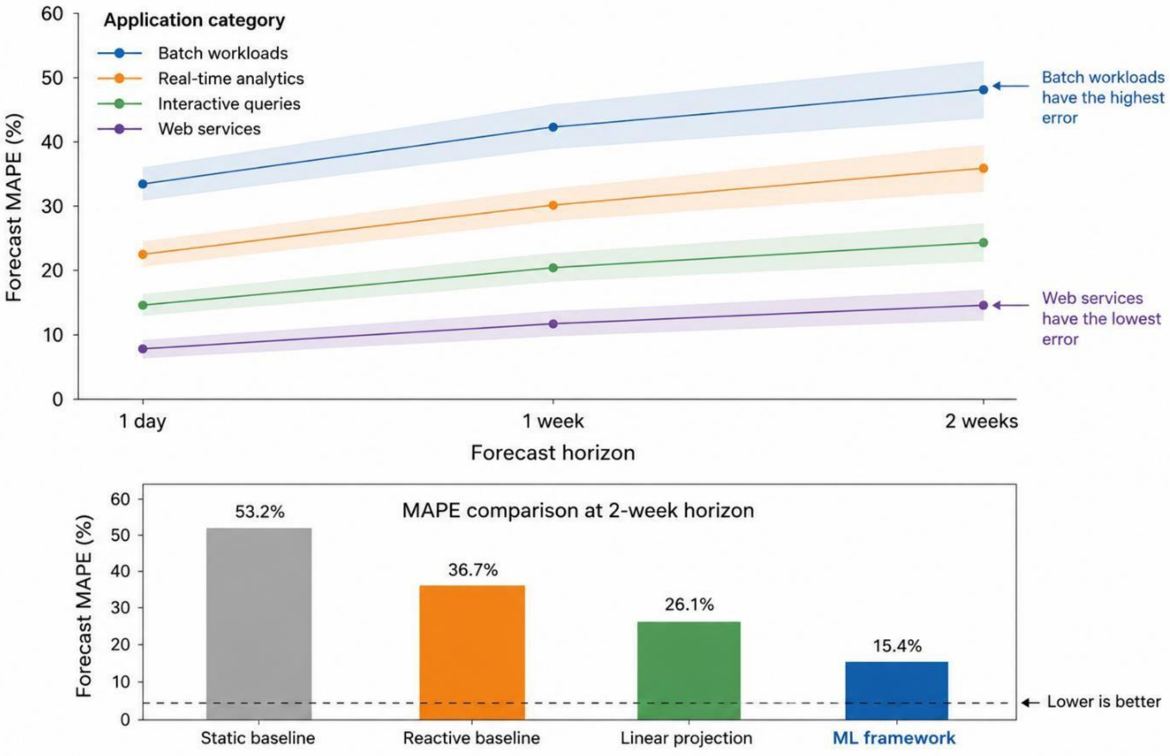

Figure 3. Forecast accuracy across horizons and application categories.

5.2 Cost and Reliability Impact

The ML framework reduced resource over-provisioning by 32% compared to static peak-based provisioning while simultaneously reducing capacity-related incidents by 58%. This dual improvement is the framework's central result: it broke the traditional trade-off between cost efficiency and reliability that forces static approaches to choose one or the other.

Table 2. Capacity planning approach comparison over 6-month evaluation.

Approach	Over-provisioning (%)	Capacity Incidents	Mean Utilization	Relative Cost
Static peak-based	54%	12	46%	1.00
Reactive threshold	31%	47	69%	0.71
Linear projection	42%	28	58%	0.84
ML predictive framework	22%	5	78%	0.68

5.3 Business Event Incorporation

Incorporating business event calendars improved forecast accuracy during high-variance periods by 27%. Month-end financial processing and marketing campaign launches, which produced demand spikes that pure time-series models missed, were anticipated accurately when business event features were included, allowing proactive capacity provisioning ahead of these known events.

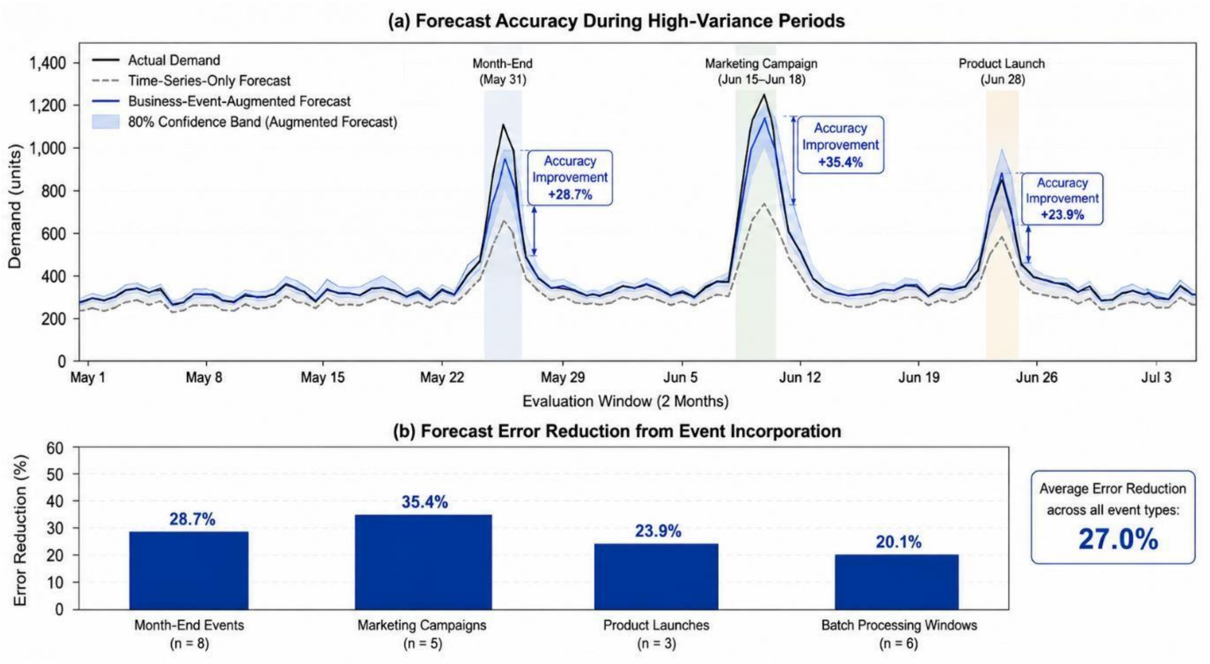

Figure 4. Impact of business event incorporation on forecast accuracy during high-variance periods.

DISCUSSION

The dual improvement in cost and reliability is the finding with the greatest practical significance. Capacity planning has traditionally been framed as a trade-off: spend more for reliability or accept risk to save money. The ML framework demonstrates that better forecasting can improve both dimensions simultaneously by provisioning the right amount at the right time rather than maintaining a static buffer sized for worst-case demand. The 32% reduction in over-provisioning represents direct cost savings, while the 58% reduction in capacity incidents represents improved reliability, and achieving both at once changes the fundamental economics of capacity management.

The value of business event incorporation highlights an important lesson about enterprise demand forecasting. Pure time-series models, however sophisticated, cannot anticipate events that have no precedent in the historical pattern, such as a new marketing campaign or a one-time data migration. By incorporating knowledge of scheduled business events, the framework anticipates demand that would otherwise appear as an unpredictable spike. This argues for close integration between capacity planning systems and the business planning processes that generate demand-driving events.

Forecast uncertainty deserves careful handling. The framework provisions to a confidence bound rather than a point forecast, trading some cost efficiency for reliability assurance. The appropriate confidence level depends on the criticality of each application, and a one-size-fits-all service level would be suboptimal. Critical applications warrant higher confidence bounds and thus more buffer, while tolerant applications can run closer to the forecast. This per-application tuning is an important practical consideration.

The cold-start problem for new applications remains a genuine limitation. New applications lack the historical telemetry the framework relies upon, so they must initially fall back to conventional provisioning until sufficient history accumulates. Transfer learning from similar applications in the portfolio could mitigate this, and represents a promising direction for future development. Model maintenance is also an ongoing concern, as workloads evolve and models trained on old patterns gradually lose accuracy, requiring periodic retraining and drift monitoring.

DOI: 10.46121/pspc.52.3.4

CONCLUSION

Predictive capacity planning using machine learning offers a compelling improvement over the static, reactive, and heuristic approaches that have long dominated enterprise capacity management. The framework evaluated here reduced over-provisioning by 32% while reducing capacity-related incidents by 58%, breaking the traditional trade-off between cost and reliability. Forecast accuracy at a two-week horizon was strong, and the incorporation of business event knowledge substantially improved accuracy during the high-variance periods that cause the most operational pain. Realizing these benefits requires more than deploying forecasting algorithms; it demands an organizational shift from reactive capacity firefighting to proactive planning, close integration with business event calendars, thoughtful handling of forecast uncertainty, and ongoing attention to model maintenance and the cold-start challenge for new applications. Organizations that make this shift will find that capacity planning transforms from a persistent source of cost and risk into a well-managed, data-driven discipline.

REFERENCES

1. Mayank Atreya, Navin Chhibber, Harvendra Singh, Explainable Machine Learning For Dynamic Pricing In Fast-Changing Retail Environments, 2022/4/9, Journal ,Available at SSRN 6011354, https://scholar.google.com/citations?view_op=view_citation&hl=en&user=fyViF1UAAAAJ&citation_for_view=fyViF1UAAAAJ:LkGwnXOMwfcC.
2. Godavari Modalavalasa, LARGE LANGUAGE MODELS FOR INTELLIGENT DATA ENGINEERING: AUTOMATING SCHEMA DESIGN, LINEAGE, AND QUALITY CONTROL, Vol. 50 No. 2 (2022): April-June 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/183 , DOI: https://doi.org/10.46121/pspc.50.2.4
3. Godavari Modalavalasa, FEDERATED LEARNING FOR ENTERPRISE CLOUD DATA ENGINEERING: ARCHITECTURE, SECURITY, AND GOVERNANCE CHALLENGES, Vol. 51 No. 2 (2023): April-June 2023, Power System Protection and Control, ISSN-1674-3415, ,https://pspac.info/index.php/dlbh/article/view/184, DOI: https://doi.org/10.46121/pspc.51.2.5
4. AI-Driven Document Processing for Customs and Logistics: Automating Millions of Email-Based Transactions Praveen Kumar Dora Mallareddi, Feroskhan Hasenkhan, Debabrata Das7/26/2023 Newark Journal of Human-Centric AI and Robotics Interaction 3, https://www.njhcair.org/index.php/publication/article/view/13
5. Naresh Lokiny. (2022). Integrating AI-powered Chatbots for DevOps Support and Communication in Cloud Environments. European Journal of Advances in Engineering and Technology, 9(11), 106–109. https://doi.org/10.5281/zenodo.13325989
6. Naresh Lokiny, (2021), "Disaster Recovery and Business Continuity Planning in DevOps Cloud with AI", International Journal of Science and Research (IJSR), 10(3), 2024-2027. https://dx.doi.org/10.21275/SR24724151733, https://www.ijsr.net/getabstract.php?paperid=SR24724151733
7. Naresh Lokiny, & Ranganath Nandanampati. (2020). DevSecOps: Integrating Security into DevOps with AI in Cloud. Journal of Scientific and Engineering Research, 7(10), 239–242. https://doi.org/10.5281/zenodo.13348695
8. Naresh Lokiny, & Pradip Reddy. (2021). Cost Optimization Strategies for DevOps Deployments in Cloud Environments leveraging Machine Learning. European Journal of Advances in Engineering and Technology, 8(3), 69–72. https://doi.org/10.5281/zenodo.13325845
9. Jayanth Para, AI Based Cloud Computation Observational Method & Process, Vol. 51 No. 4 (2023): October-December 2023, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/171 , DOI: https://doi.org/10.46121/pspc.51.4.3
10. Jobayar Alom, Ahsan Ahmed. (2023). Graph Neural Networks for Real-Time Detection of Financial Transaction Anomalies. Acta Scientiae, 24(5), 82–90. https://doi.org/10.22178/acta.24.5.6
11. **Kaleshwar Aryasomayajula**, PREVENTING BIAS IN AI-BASED DECISION-MAKING: ANALYZING TECHNIQUES TO REMOVE UNFAIR PREJUDICE IN ALGORITHMIC OUTCOMES, Vol. 50 No. 2 (2022):

April-June 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/294

12. Kaleshwar Aryasomayajula, MODERN DEVELOPMENT PRACTICES: INVESTIGATIONS INTO SERVERLESS COMPUTING, LOW-CODE/NO-CODE PLATFORMS, AND DEVELOPER PRODUCTIVITY IN REMOTE ENVIRONMENTS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/364
13. **Vikas Suresh Bagora**, KERNEL-LEVEL PERFORMANCE OPTIMIZATION FOR CARRIER-GRADE BROADBAND AND HIGH-SPEED NETWORKING SYSTEMS, Vol. 47 No. 1 (2019), Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/359
14. **Vikas Suresh Bagora,** SCALABLE 128G FIBRE CHANNEL AND ETHERNET CONVERGENCE FOR NEXT-GENERATION DATA CENTER NETWORKS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/362
15. **Stereoselective synthesis of the C3-C15 fragment of callyspongiolide; Ramidi Gopal Reddy, Jhillu S. Yadav and Debendra K. Mohapatra. (Tetrahedron Letters.** 2018, 59, 3579-3582). https://doi.org/10.1016/j.tetlet.2018.08.041, https://www.sciencedirect.com/science/article/pii/S0040403918310426
16. **Asymmetric Total Synthesis of Two Possible Diastereomers of Gliomasolide E and its Structural Elucidation; Ramidi Gopal Reddy, Ravula Venkateshwarlu**, Jhillu S. Yadav and Debendra K. Mohapatra. (J. Org. Chem. 2017, 82(2), 1053-1063). https://doi.org/10.1021/acs.joc.6b02611 https://pubs.acs.org/doi/abs/10.1021/acs.joc.6b02611
17. **Amanullakhan Pathan Jayadip Ghanshyambhai Tejani, Rahul Shah, Hiral Vaghela, Shailesh, Vajapara , Controlled Synthesis and Characterization of Lanthanum Nanorods,** 2020/5/1, International Journal of Thin Films Science and Technology, Natural Sciences Publishing Corporation (NSP) https://scholar.google.com/citations?view_op=view_citation&hl=en&user=efbNMkwAAAAJ&citation_for_view=efbNMkwAAAAJ:M3NEmzRMIkIC
18. **Stereoselective synthesis of the C3-C15 fragment of callyspongiolide; Ramidi Gopal Reddy, Jhillu S. Yadav and Debendra K. Mohapatra. (Tetrahedron Letters**. 2018, 59, 3579-3582). https://doi.org/10.1016/j.tetlet.2018.08.041, https://www.sciencedirect.com/science/article/pii/S0040403918310426
19. **Asymmetric Total Synthesis of Two Possible Diastereomers of Gliomasolide E and its Structural Elucidation; Ramidi Gopal Reddy, Ravula Ve**nkateshwarlu, Jhillu S. Yadav and Debendra K. Mohapatra. (J. Org. Chem. 2017, 82(2), 1053-1063). https://doi.org/10.1021/acs.joc.6b02611, , https://pubs.acs.org/doi/abs/10.1021/acs.joc.6b02611
20. **Kaleshwar Aryasomayajula**, PREVENTING BIAS IN AI-BASED DECISION-MAKING: ANALYZING TECHNIQUES TO REMOVE UNFAIR PREJUDICE IN ALGORITHMIC OUTCOMES, Vol. 50 No. 2 (2022): April-June 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/294
21. **Kaleshwar Aryasomayajula, MODERN DEVELOPMENT PRACTICES: INVESTIGATIONS INTO SERVERLESS** COMPUTING, LOW-CODE/NO-CODE PLATFORMS, AND DEVELOPER PRODUCTIVITY IN REMOTE ENVIRONMENTS, Vol. 50 No. 4 (2022): October-December 2022, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/364
22. **Kaleshwar Aryasomayajula**, Micro services: "A Comparative Analysis of Monolithic vs. Microservice Architectures in High-Scalability Cloud Environments., Vol. 51 No. 2 (2023): **April-June 2023** , Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/374

23. **Controlled Synthesis and Characterization of Lanthanum Nanorods**, **Author(s), Jayadeep Tejani, Rahul Shah, Hiral Vaghela, Shailesh Vajapara, Amanullakhan Pathan**, doi:10.18576/ijtfst/090205, URL: https://www.naturalspublishing.com/Article.asp?ArtcID=20705, **May-2020**
24. **Sudipkumar Ghanvat , Aishwarya Badlani, Shreya Sandeep Joshi,** Marketing Effectiveness in the Post-Cookie Era: A Comparative Analysis of First Party and Zero-Party Data Strategies on Campaign Performance and Customer Equity, **Vol. 31 No. 4 (2023):** JOCAAA , https://www.eudoxuspress.com/index.php/pub/article/view/3940
25. **Chander Vijay S Sanbhi,** BAYESIAN UPDATING OF RAM MODELS FOR DYNAMIC AVAILABILITY FORECASTING UNDER REALISTIC DOWNTIME CONSTRAINTS**, Vol. 51 No. 3 (2023): July-**September 2023, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/381
26. **Chander Vijay S Sanbhi,** HYBRID RELIABILITY MODELLING FOR ROTATING EQUIPMENT: PHYSICS-INFORMED LEARNING WITH SPARSE FAILURE DATA, Vol. 51 No. 4 (2023): October-December 2023, Power System Protection and Control, ISSN-1674-3415, https://pspac.info/index.php/dlbh/article/view/382
27. A. Srivastava, "AI-assisted cloud migration of healthcare claims data from legacy systems: A metadata-driven framework," Int. J. Commun. Networks Inf. Secur., vol. 14, no. 3, Nov. 2022. [Online]. Available: https://ijcnis.org/index.php/ijcnis/article/view/8682
28. A. Srivastava, "Enhancing provider and claims data accuracy using artificial intelligence on cloud-native data platforms," J. Comput. Anal. Appl., vol. 31, no. 4, Nov. 2023. [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/4347/3189
29. A. Srivastava, "Optimizing large-scale data migration for cloud-native architectures," J. Comput. Anal. Appl., vol. 31, no. 2, pp. 652–662, May 2023. [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/4603