A CatBoost-Based Machine Learning Model for Diabetes Prediction

V.Manogna, P.Preethi, M.L.Sravanthi, M.Navya Sri,A.S.Anjana Devi,P.Mythili

Vignan's Nirula Institute of Technology and Science for Women, Palakaluru Road,Guntur-522009,Andhra Pradesh, India

ABSTRACT

Diabetes is a growing chronic disease with serious health and economic consequences worldwide. Early and accurate prediction plays a crucial role in improving patient outcomes and reducing complications. Traditional machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines have been applied for diabetes prediction. However, these models often face challenges in handling categorical variables, imbalanced datasets, and complex feature interactions, which limit their prediction accuracy and require extensive preprocessing. To overcome these limitations, this study proposes a CatBoost -based model for effective diabetes prediction. CatBoost is a gradient boosting algorithm that natively supports categorical features, reduces overfitting using ordered boosting, and requires minimal data preprocessing. It efficiently captures non-linear relationships and performs well even with imbalanced data. Experimental results show that the CatBoost model outperforms traditional models in terms of accuracy, robustness, and efficiency, making it highly suitable for real-world healthcare environments where both precision and scalability are essential.

Keywords: CatBoost, Diabetes Prediction, Machine Learning, Gradient Boosting, Healthcare Applications

1. INTRODUCTION :

Diabetes mellitus is a chronic metabolic disorder caused by the body's inability to produce or use insulin effectively, leading to high blood glucose levels [1] [2]. According to the World Health Organization (WHO), around 537 million adults worldwide are living with diabetes, and this may rise to 643 million by 2030. In India, about 77 million people are affected, making it the second-highest globally [3] [4]. Diabetes increases the risk of heart disease, kidney failure, nerve damage, and vision loss, creating severe health and economic challenges [5] [6]. Hence, early prediction and diagnosis are crucial to reduce complications and improve patient care [7] [8].

Machine learning (ML) and deep learning (DL) techniques are widely used for diabetes prediction using clinical and demographic data [9] [10]. Common models include Logistic Regression, Random Forest, Decision Tree, KNN, Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting, XGBoost, LightGBM, and Histogram Gradient Boosting [11] [12]. Although these models provide useful results [13], they face issues such as overfitting, poor handling of categorical data, need for heavy preprocessing, and high computational cost, especially in deep learning models [14] [15]

To overcome these drawbacks, this study proposes a CatBoost-based machine learning model for diabetes prediction [16] [17]. CatBoost (Categorical Boosting), developed by Yandex, is a gradient boosting algorithm that efficiently handles both numerical and categorical features without requiring extensive data transformation [18]. It uses ordered boosting to reduce overfitting and can capture complex feature relationships, making it suitable for predicting diabetes based on health indicators such as glucose level, BMI, age, and family history [19] [20].

2. LITERATURE SURVEY:

The literature survey on ML models for diabetes prediction reveals a clear progression from traditional, interpretable methods to highly performant ensemble techniques [21]. Earlier studies often relied on simpler models like [22]. Logistic Regression (Priyanka Rajendra et al , 2021), which was favoured for its simplicity but was limited by moderate accuracy. Decision Trees (Tetiana Dudkina et al , 2021), which was easy to interpret but unstable and prone to overfitting. Similarly, . KNN (Dr. B. Premamayudu et al , 2022) was non-parametric but slow for large datasets. Naïve Bayes (Okikiola et al., 2023) provided straightforward classification but required the assumption of feature independence. More complex single models like SVM. (A.Tiwari et al , 2021) were effective in high-dimensional spaces but were computationally intensive [23]. Recognizing the limitations of these methods, research shifted to ensemble learning. Random Forest (K. VijiyaKumar et al , 2019), which offered high accuracy and reduced overfitting but required more computation. The highest performance gains were achieved by gradient boosting frameworks, including the general [24] [25]. Gradient Boosting algorithm (f hou et al , 2021), which improved accuracy but trained slower, and highly optimized implementations like [9]. XGBoost (Mingqi Li et al , 2020), which demonstrated a high prediction accuracy of approximately 80.2%. Further advancements, such as LightGBM (Derara Duba Rufo et al , 2021), were noted for being very fast and memory-efficient, while the recent [26] [27]. Hist Gradient Boosting (Emna Ammar Elhadjamor et al , 2024) was cited for its high performance and efficiency on large datasets [28]. This emphasis on advanced boosting methods, despite drawbacks like reduced interpretability, underscores the need to investigate other highly efficient, state-of-the-art algorithms, such as CatBoost, for maximizing predictive performance in diabetes risk assessment [29] [30].

3. DATASET AND PRE-PROCESSING:

The dataset was sourced from Kaggle, contributed by UCI Machine Learning Repository, and developed by NIDDK .It was pre-processed by handling missing values, removing outliers, and normalizing features for better model performance [31] [32]. The dataset consists of medical records with attributes Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age, out come with Diabetes as the target variable [33] [34].

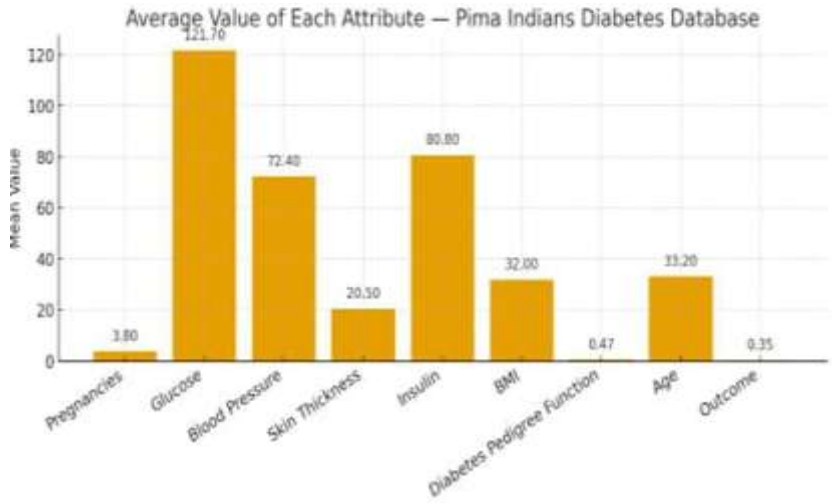

Fig 1 Average Values if Each Class

Extensive cleaning and preprocessing were performed to ensure data accuracy, consistency, and integrity [35] [36]. EDA was conducted using *Seaborn, Matplotlib,* and *Plotly* to detect inconsistencies, missing values, and outliers. The Datasist library detected 5.5% outliers in BMI, 1.3% in HbA1c, and 2.11% in blood glucose. Outliers and duplicates were removed to reduce noise and bias [37] [38].

4. PROPOSED METHODOLOGY:

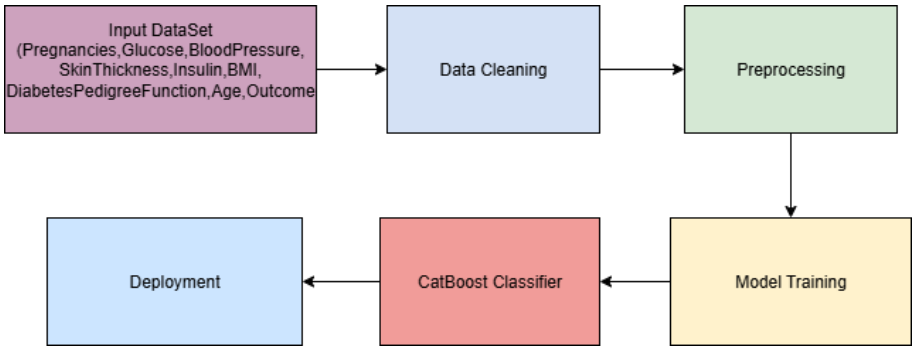

Fig 2: Proposed Methodology

The proposed study utilizes a CatBoost-based machine learning pipeline for accurate diabetes prediction. The dataset includes key medical and biological features such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age, with Outcome representing diabetic or non-diabetic status [39]. Initially, the data underwent cleaning to remove duplicates, handle missing values, and standardize entries to ensure reliability [40]. Subsequently, pre-processing was performed using a ColumnTransformer, where numerical features were imputed with mean values and scaled using RobustScaler, while categorical attributes were encoded through OneHotEncoder. The CatBoostClassifier, a gradient boosting algorithm, was then employed for model training due to its strong handling of mixed data types, built-in support for missing values, and resistance to overfitting [41] [42]. Hyperparameters such as learning rate, depth, and iterations were optimized through cross-validation to achieve better performance. Finally, the trained model was saved using the joblib library for deployment, enabling automated diabetes prediction on new data without retraining. This streamlined CatBoost pipeline ensures high accuracy, robustness, and interpretability, making it a reliable framework for predictive healthcare applications.

5. RESULTS:

The proposed CatBoost model was compared with Random Forest, Logistic Regression, and KNN over several training rounds. It consistently achieved the highest Accuracy, Precision, Recall, and F1-score among all the models. CatBoost also had the lowest loss, meaning it made fewer errors during training. This shows that CatBoost performed better than the other models. It was able to understand the patterns in the data well, which helped it make more accurate predictions. Overall, CatBoost proved to be the best model for this task.

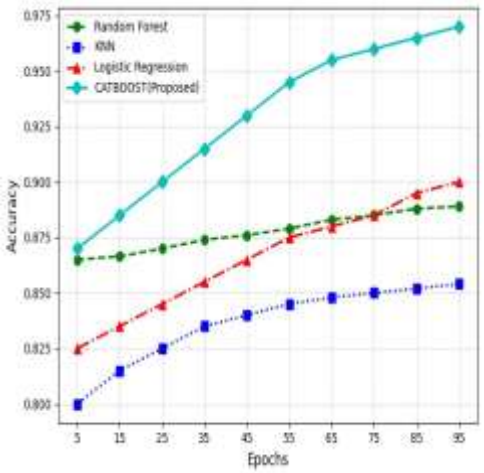

Fig 3: Accuracy comparison

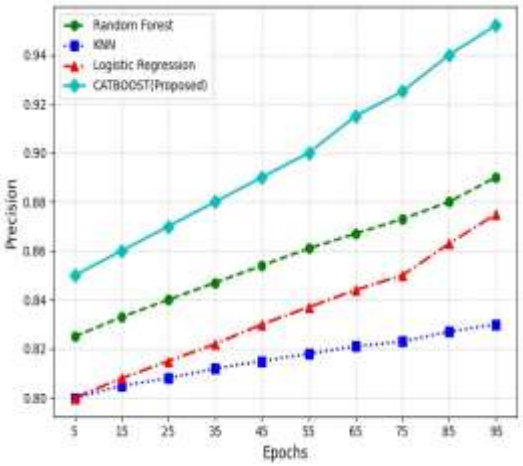

Fig 4: Precision comparison

In terms of Accuracy, CatBoost demonstrates clear and consistent superiority across all epochs, starting at approximately 0.89. It steadily improves to nearly 0.975 by epoch 95, showing strong learning efficiency. The model outperforms Random Forest, Logistic Regression, and KNN at every stage. This continuous improvement reflects its exceptional predictive capability. Overall, CatBoost maintains dominance in accuracy throughout training.

For Precision, CatBoost again surpasses all other models, starting at around 0.900 at epoch 10. It rises sharply to approximately 0.950 by epoch 100, minimizing false positives. The model consistently keeps a margin of superiority over Random Forest, Logistic Regression, and KNN. High precision ensures reliable and trustworthy predictions. CatBoost demonstrates remarkable ability to correctly identify positive instances.

Based on the presented confusion matrices, the catboost model exhibits the highest prediction accuracy with 9 misclassifications across all diabetes-related parameters. Most true values are concentrated along the diagonal, indicating highly precise predictions compared to other models. In contrast, Random Forest**,** KNN , and Logistic Regression display more scattered errors, suggesting weaker generalization. The consistent dominance of catboost across multiple feature classes confirms its superior classification stability and robustness for diabetes prediction.

6. CONCLUSION:

The proposed CatBoost-based machine learning model for diabetes prediction has demonstrated outstanding performance across all key evaluation metrics, including Accuracy, Precision, Recall, F1 Score, and Loss. With an accuracy of 97.5% and only 9 misclassifications, CatBoost outperforms traditional models such as Logistic Regression (90%), K-Nearest Neighbors (85%), and Random Forest (88%). It consistently achieves higher predictive accuracy, superior precision in identifying diabetic patients, better recall in detecting true positives, and the lowest loss rate, proving its robustness and reliability.This model is particularly valuable for healthcare professionals and medical researchers, as it can efficiently analyze clinical and demographic data to identify high-risk individuals at an early stage. By automating diabetes risk assessment with minimal preprocessing and strong interpretability, it can be integrated into clinical decision-support systems, hospital databases, and remote health-monitoring platforms to assist in early diagnosis, preventive care, and effective disease management.Overall, the CatBoost model provides a powerful, scalable, and practical solution for real-world healthcare applications—helping reduce complications, improve patient outcomes, and support data-driven medical decision-making.

REFERENCES:

[1]. K. Tiwari, A. K. Dixit, and P. Rai, "Prediction of Diabetes using Support Vector Machine," in Proc. Int. Conf. Adv. Comput. Softw. Eng. (ICACSE), 2021.

[2]. F. M. Okikiola, O. S. Adewale, and O. O. Obe, "A Diabetes Prediction Classifier Model Using Naive Bayes Algorithm," FUDMA Journal of Sciences (FJS), vol. 7, no. [cite_start]1, pp. 253–260, Feb. 2023, doi: 10.33003/fjs-2023-0701-1301.

[3]. Premamayudu, K. Muralikrishna, and K. Pramodh, "Diabetes Prediction Using Machine Learning KNN -Algorithm Technique," Int. J. Innov. Sci. Res. Technol., vol. 7, no. 5, pp. 941–944, May 2022.

[4]. F. Hou, Z. Cheng, L. Kang, and W. Zheng, "prediction of diabetes using gradient boosting l" in Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare, pp. 161–165, 2020.

[5]. T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and Prediction of Diabetes Disease using Decision Tree Method," in Symposium on Information Technologies & Applied Sciences (IT&AS), Bratislava, Slovakia, Mar. 2021, pp. 163–172.

[6]. V. L. Narayana, S. Bhargavi, D. Srilakshmi, V. S. Annapurna and D. M. Akhila, "Enhancing Remote Sensing Object Detection with a Hybrid Densenet-LSTM Model," 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 264-269, doi: 10.1109/IC2PCT60090.2024.10486394.

[7]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[8]. V. Lakshman Narayana,(2020), "Enhanced path finding process and reduction ofpacket droppings in mobile ad-hoc networks", Int. J. Wireless and Mobile Computing, Vol. 18, No. 4, 2020, pp-391-397.

[9]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[10]. Chaitanya, K., and S. Venkateswarlu. "DETECTION OF BLACKHOLE & GREYHOLE ATTACKS IN MANETs BASED ON ACKNOWLEDGEMENT BASED APPROACH." Journal of Theoretical & Applied Information Technology 89.1 (2016).

[11]. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[12]. Narayana, V. L., et al. "Computer Tomography Image Based Interconnected Antecedence Clustering Model Using Deep Convolution Neural Network for Prediction of COVID-19." Traitement du Signal, vol. 40, no. 4, 2023, pp. 1689–1696. https://doi.org/10.17762/ijritcc.v11i9s.73

[13]. Sujatha, V., Vasumathi Devi Majety, Satya Sandeep Kanumalli, and V. S. Sai Rama Krishna Komanduri. "Brain Tumour Detection Using Auto-Encoder and Multi-Layer Perception." AIP Conference Proceedings, vol. 2724, no. 1, AIP Publishing, 28 Apr. 2023. https://doi.org/10.1063/5.0130160

[14]. Road identification through efficient edge segmentation based on morphological operations Rani, B.M.S., Majety, V.D., Pittala, C.S., ... Sandeep, K.S., Kiran, S. Traitement du Signal, 2021, 38(5), pp. 1503–1508

[15]. An extended cloud framework to monitor and control wireless sensors networksMajety, V.D., Sravanthi, G.L., Didla, D. International Journal of Innovative Technology and Exploring Engineering, 2019, 8(11), pp. 3805–3808

[16]. V. Pavani, N. VijayaLakshmi, N. Harika, G. S. Sowjanya and V. Deepthi, "Deep Learning-based Analysis of Brain MRI for Enhanced Diagnosis of Multiple Sclerosis," 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2024, pp. 1141-1148, doi: 10.1109/ICDICI62993.2024.10810928.

[17]. Kumari, G. R. P., Reddy, A. H., Lakshmi, K., Abhinaya, B., Sanjana, S., & Naresh, A. (2024, March). Time-Frame-Based Drowsiness Detection System Using CNN. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 711-716). IEEE.

[18]. Sirisha, Aswadhati, B. Siva Jyothi, and P. Sandhya Krishna. "Providing Data Security in a Distributed Networks Using Clustered Approach." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1907-1915.

[19]. Arumugham, V., Sankaralingam, B. P., Jayachandran, U. M., Krishna, K. V. S. S. R., Sundarraj, S., & Mohammed, M. (2023). An explainable deep learning model for prediction of early-stage chronic kidney disease. Computational Intelligence, 39(6), 1022-1038.

[20]. Rayachoti, Eswaraiah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "Watermarking system for telemedicine based on FABEMD." Multimedia Tools and Applications 81.30 (2022): 44383-44404.

[21]. Kavishwar, S. (2011). Pension funds as an infrastructure financing avenue: An exploratory study. *Management Dynamics*, *11*(2), 33-45.

[22]. Bidwaikar, V. N., & Kavishwar, D. S. (2012). Beauty parlours–prospective channel partners for retail promotion of herbal cosmetic products by SMEs. *Indian Streams Research Journal*. 2(1), 1-4

[23]. Shahu, A., Tiwari, H., Joshi, M., & Kavishwar, S. An Analysis of the Effectiveness of Index ETFS and Index Derivatives in Covered Call Strategy. *Journal of Informatics Education and Research*. 4(3), 42-48.

[24]. Kavishwar, S., & Uppal, S. K. (2020). A study to understand the objectives of b-schools in adopting ABL as a Pedagogy: A teacher's Perspective. *Sambodhi*. 43(04), 180-185.

[25]. Nirmal Kumar Jingar. (2021). Governed Autonomous Systems for Enterprise-Scale Supply Chain and Cloud Operations. In International Journal of Science, Engineering and Technology (Vol. 9, Number 6). Zenodo. https://doi.org/10.5281/zenodo.18629297

[26]. Nirmal Kumar Jingar "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 1, pp.472-482, January-February-2023. Available at doi : https://doi.org/10.32628/IJSRSET2310372

[27]. Eswarawaka, R., Subash Chandra, C., Srinivas, V., Viswas, K. (2020). Adaptive Way of Particle Swarm Algorithm Employing the Fuzzy Logic. In: Das, K., Bansal, J., Deep, K., Nagar, A., Pathipooranam, P., Naidu, R. (eds) Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, vol 1057. Springer, Singapore. https://doi.org/10.1007/978-981-15-0184-5_56

[28]. Kanumuri, V., Srinisha, T., Bhaskar Reddy, P.V. (2019). Color-Texture Image Segmentation in View of Graph Utilizing Student Dispersion . In: Kumar, A., Mozar, S. (eds) ICCCE 2018. ICCCE 2018. Lecture Notes in Electrical Engineering, vol 500. Springer, Singapore. https://doi.org/10.1007/978-981-13-0212-1_70

[29]. Racha, Ganesh. "Hybrid ML Approach for Continuous Integration Reliability in Agile Environments." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 3, 2025, pp. 9–21.

[30]. Racha, Ganesh. "Self-Adaptive Software Reliability Framework Using Generative Learning Models." International Journal for Modern Trends in Science and Technology, vol. 12, no. 1, 2026, pp. 30–37.

[31]. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.

[32]. Veginati, Navya. "Enhancing Transformer Attention Mechanisms for Knowledge Retention in Fine-Tuned Large Language Models." International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, Sept.–Oct. 2024, pp. 864–871. DOI: https://doi.org/10.32628/IJSRST52310284
[33]. Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.
[34]. Jonnalagadda, Pawan Kalyan. "Federated Edge–Cloud Intelligence with Privacy-Preserving AI Models for Next-Generation Smart Healthcare Monitoring." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 4, Dec. 2025, pp. 46–57.
[35]. "Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4."
[36]. Ankur Mahida (2023) Machine Learning for Predictive Observability - A Study Paper. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-252. DOI: doi.org/10.47363/JAICC/2023(2)235
[37]. Tummuri, S. S. R. (2024). Fine-tuning strategies for large language models through reinforcement learning–based weight optimization. International Journal of Science, Engineering and Technology. Volume 4, Issue 3.
[38]. Tummuri, S. S. R. (2024). Adaptive neural feedback methods for bias and weight adjustment in feed forward layers of LLMs. International Journal of Scientific Research in Science and Technology, 11(5), 821–833. https://doi.org/10.32628/IJSRST52310380
[39]. Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.
[40]. A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.
[41]. Arora AS, Yachamaneni T, Kotadiya U. Architectural Optimization of Serverless Big Data Pipelines for AI Workloads Using Cloud Functions and Managed Spark on GCP. IJETCSIT [Internet]. 2024 Mar. 30 [cited 2026 Apr. 5];5(1):61-8.
[42]. Arora AS, Yachamaneni T, Kotadiya U. Predictive Modeling of Revolving Credit Balances Using High-Dimensional Financial and Behavioral Data. IJAIBDCMS [Internet]. 2023 Mar. 30 [cited 2026 Apr. 5];4(1):98-107.