A Comprehensive Study on Machine Learning Approaches for IoT Data Classification

Hima Bindu Paka[1*], Gaddam Nandini[2], Bodhireddy Shruthi[2], Damera Sanjay[2], Gaddala Manideep Babu[2]

[1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: himabindu.paka@gmail.com

ABSTRACT

The exponential growth of the Internet of Things (IoT) has led to vast amounts of data being generated, requiring efficient classification methods for real-time analysis and decision-making. Traditional classification techniques often struggle with high-dimensional, imbalanced IoT data, affecting predictive performance. This study presents a comprehensive machine learning-based approach for IoT data classification, incorporating Exploratory Data Analysis (EDA), data preprocessing, and model evaluation. The dataset undergoes rigorous cleaning, including handling missing values, duplicate removal, label encoding for categorical variables, and data balancing through resampling. Various machine learning models are implemented, including Logistic Regression and CatBoost Classifier, to assess classification performance. EDA is conducted using statistical summaries, null value checks, duplicate detection, and feature distribution visualization through count plots. The dataset is split into training and testing sets (80-20 ratio), and performance metrics such as accuracy, precision, recall, and F1-score are calculated. Logistic Regression serves as a baseline model, achieving an accuracy of 20.31%, with low precision and recall, indicating poor classification performance. In contrast, the CatBoost Classifier demonstrates superior performance, achieving a perfect accuracy of 100%, precision of 100%, recall of 100%, and F1-score of 100%. Comparative analysis using confusion matrices highlights the ability of CatBoost to effectively classify IoT device categories without misclassification. The results emphasize the advantages of using advanced gradient-boosting algorithms like CatBoost over traditional models for IoT data classification. The study provides a robust framework for real-time IoT device classification, ensuring higher accuracy and efficiency in handling complex, high-frequency IoT datasets.

Keywords: Internet of Things, Data classification, Decision making, CatBoost classifier.

1. INTRODUCTION

The Internet of Things (IoT) has seen significant evolution from simple sensor networks to a complex ecosystem of interconnected devices that generate vast amounts of data. The volume of IoT data is growing exponentially, with projections estimating it will reach 79.4 zettabytes by 2025, up from 13.6 zettabytes in 2019, fueled by the increasing adoption of over 30 billion IoT devices by the same year. This explosion of data has made managing, processing, and classifying it a critical challenge. IoT data classification, which involves categorizing data from various devices into meaningful classes, is essential for applications ranging from smart homes to industrial monitoring. Traditional classification methods, such as rule-based systems and simple algorithms like decision trees and linear classifiers, struggle to handle the complexity and volume of modern IoT data. These methods are prone to issues such as overfitting, limited generalization, and an inability to capture intricate patterns, making them ill-suited for the dynamic, high-dimensional, and noisy nature of IoT data. To address these limitations, there is a growing need for advanced machine learning approaches capable of efficiently managing heterogeneous and dynamic IoT data, thus improving the accuracy and reliability of IoT systems. By

adopting deep learning and ensemble methods, it is possible to enhance the classification process, reduce the need for manual feature engineering, and improve performance, especially in applications requiring real-time processing and high accuracy, such as autonomous vehicles and smart grid management. These advanced techniques can also make IoT systems more adaptive and scalable, helping them evolve with the data and environments in which they operate. The motivation for transitioning from traditional manual classification approaches to machine learning-driven solutions lies in the need for automation, higher accuracy, scalability, and real-time decision-making. Manual approaches are not feasible for the massive and continuous data streams generated by IoT devices, particularly in industries requiring immediate responses to sensor data, such as predictive maintenance and health monitoring. Automated machine learning models offer the ability to classify data accurately

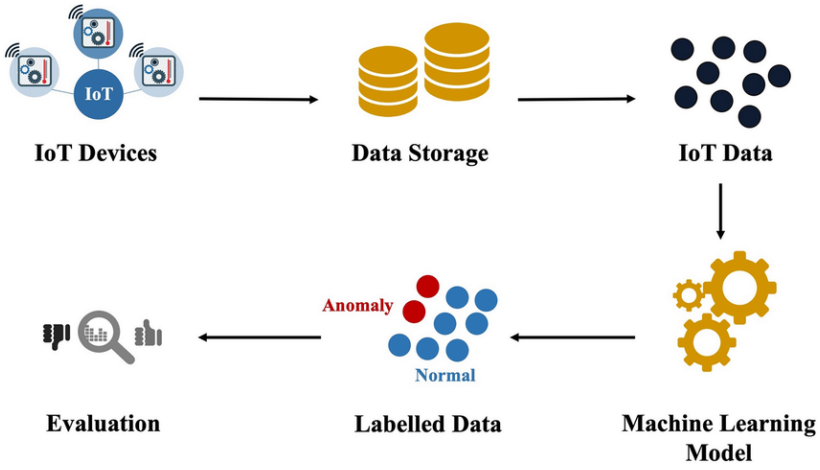


Figure 1: Machine Learning on IoT Data.

and consistently, even in high-dimensional, noisy environments, and enable real-time decision-making. The adoption of machine learning for IoT data classification is crucial to unlocking the full potential of IoT applications. Key performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of machine learning models, including LightGBM and Random Forest classifiers. These metrics assess the models' ability to correctly predict labels, avoid false alarms, identify critical faults, and balance overall reliability, making them essential for real-world IoT applications in areas like smart cities, healthcare, industrial automation, and agriculture, where precise, real-time classification can have significant operational benefits.

2. LITERATURE SURVEY

Mahdavinejad et al. [1] explored the architecture and opportunities in big IoT data analytics, addressing the key challenges and proposing solutions for handling vast amounts of data generated by IoT devices. Their work emphasized the importance of a scalable and efficient architecture that can support real-time analytics, enabling more intelligent and responsive IoT systems. By identifying open research challenges, they set the stage for future innovations in IoT data analytics. Zahoor and Mir [2] conducted a comprehensive survey on the application of machine learning techniques in IoT data analysis. Their research focused on how various machine learning models can be utilized to process and analyze the enormous data generated by IoT devices, providing insights into the advantages and limitations of different approaches. They highlighted the potential of machine learning to enhance the intelligence and decision-making capabilities of IoT systems. Patel et al. [3] presented a detailed survey on resource management in pervasive IoT environments, examining the strategies for optimizing the use of computational, storage, and network resources in IoT systems. Their study covered various resource

management techniques, including load balancing, energy efficiency, and quality of service, offering a comprehensive overview of how to efficiently manage resources in large-scale IoT deployments.

Vukobratovic [4] discussed the use of intelligent edge computing for IoT analytics, focusing on how edge devices can be leveraged to perform data processing closer to the data source. This approach reduces latency and bandwidth usage while improving the responsiveness of IoT applications. The study highlighted the potential of edge computing to enhance the scalability and efficiency of IoT systems, especially in scenarios requiring real-time decision-making. Guinard et al. [5] proposed the CONDENSE architecture, a reconfigurable knowledge acquisition framework for 5G IoT systems. Their work emphasized the need for flexible and adaptive systems that can handle the dynamic and heterogeneous nature of IoT data in 5G networks. By providing a scalable and modular architecture, their research aimed to facilitate the integration of diverse IoT devices and services into future 5G ecosystems. Guinard et al. [6] explored interactions with SOA-based IoT systems, focusing on the discovery, query, selection, and on-demand provisioning of web services. Their research addressed the challenges of integrating IoT devices into service-oriented architectures, proposing solutions for efficient service management in IoT environments. This work provided a foundation for developing more interoperable and flexible IoT systems that can easily adapt to changing user requirements. Likas et al. [7] introduced the global k-means clustering algorithm, which is designed to overcome the limitations of traditional k-means algorithms in handling large datasets. Their approach iteratively adds one cluster center at a time, ensuring that the final clustering solution is globally optimal. This method has been widely adopted in various applications, including IoT data analysis, where it helps in efficiently organizing and interpreting large-scale data. Singh and Reddy [8] provided a survey on big data analytics platforms, highlighting the key features, capabilities, and limitations of existing platforms. Their research focused on the tools and technologies used to process and analyze big data, particularly in the context of IoT. By comparing different platforms, they offered valuable insights into selecting the most appropriate tools for specific big data applications.

Coates and Ng [9] presented a study on learning feature representations using K-means clustering, demonstrating how unsupervised learning techniques can be used to extract meaningful features from large datasets. Their research has significant implications for IoT data analysis, where efficient feature extraction is crucial for improving the performance of machine learning models. Bro and Smilde [10] discussed the application of principal component analysis (PCA) in data analysis, focusing on how PCA can be used to reduce the dimensionality of large datasets while preserving the most important information. Their work provided a foundation for using PCA in IoT data analysis, where it helps in simplifying complex data and making it more manageable for further processing. Lecun et al. [11] explored gradient-based learning techniques applied to document recognition, laying the groundwork for many modern machine learning applications. Their research is highly relevant to IoT data analytics, where gradient-based methods are often used to train models on large and complex datasets, enabling more accurate and efficient data processing.

Qin et al. [12] conducted a survey on data-centric IoT, focusing on how data can be managed, processed, and utilized in IoT systems. Their work emphasized the importance of data as a central element in IoT, exploring various techniques for data storage, management, and analysis. By highlighting the challenges and opportunities in data-centric IoT, they provided a comprehensive overview of the current state and future directions in this field. Yogita and Toshniwal [13] surveyed clustering techniques for streaming data, examining how these methods can be applied to continuous data streams generated by IoT devices. Their research addressed the unique challenges of clustering streaming data, such as handling data velocity and ensuring real-time processing. They provided a detailed comparison of

different clustering techniques, offering guidance on selecting the most appropriate methods for specific IoT applications.

Ni et al. [14] proposed a hybrid method for short-term sensor data forecasting in IoT environments, combining statistical models with machine learning techniques. Their research focused on improving the accuracy and reliability of sensor data predictions, which is crucial for real-time IoT applications. By integrating different forecasting methods, they aimed to enhance the robustness and adaptability of IoT systems in dynamic environments. Ma et al. [15] analyzed smart card data to identify transit riders' travel patterns, demonstrating how IoT data can be used to gain insights into urban mobility. Their work involved the application of data mining techniques to extract meaningful patterns from large datasets, providing valuable information for transportation planning and management. This study highlighted the potential of IoT data to improve the efficiency and effectiveness of urban transit systems. Derguech et al. [16] proposed an autonomic approach to real-time predictive analytics using open data and IoT, focusing on how these technologies can be combined to create more responsive and adaptive systems. Their research emphasized the importance of real-time data processing in IoT, particularly for applications that require immediate action based on data insights. By integrating open data with IoT analytics, they provided a framework for developing more intelligent and autonomous systems.

Luss and Aspremont [17] explored the use of text classification techniques to predict abnormal returns from news articles, demonstrating how IoT data can be combined with other data sources to generate actionable insights. Their research focused on the financial domain, where timely and accurate predictions are crucial for making informed decisions. This study provided a novel application of text classification in IoT analytics, showcasing the potential of combining diverse data types for enhanced predictive capabilities. Han et al. [18] proposed a data-driven quantitative trust model for the Internet of Agricultural Things, focusing on how trust can be established and maintained in IoT systems used in agriculture. Their work addressed the unique challenges of IoT in agricultural environments, such as ensuring data reliability and security. By developing a trust model based on data analysis, they aimed to improve the adoption and effectiveness of IoT technologies in agriculture. Souza and Amazonas [19] developed an outlier detection algorithm for IoT architectures using big data processing techniques. Their research focused on identifying and handling anomalous data points in large IoT datasets, which is crucial for maintaining the accuracy and reliability of IoT systems. By leveraging big data processing capabilities, they provided a scalable solution for outlier detection in IoT environments, contributing to the overall robustness of these systems.

3. PROPOSED METHODOLOGY

Step 1: Dataset Uploading

The process begins by importing the necessary libraries such as pandas, seaborn, and various machine learning libraries for data manipulation and modeling. The IoT dataset, typically in CSV format, is then uploaded into the environment using pandas' read_csv function. An initial examination of the dataset follows, where any duplicates and missing values are checked. This step ensures the dataset is clean, free of redundancies, and prepared for the next stages of analysis and processing.

Step 2: Data Preprocessing

In the data preprocessing phase, the dataset undergoes several transformations to make it suitable for machine learning algorithms. First, any missing values in the dataset are handled, often by removing rows or filling them with relevant values depending on the context. Additionally, categorical data (non-numeric variables) are converted into numerical format using LabelEncoder from scikit-learn. This

transformation is essential because machine learning models require numerical inputs, and encoding categorical variables allows the model to interpret them appropriately for classification tasks.

Step 3: Data Resampling and Splitting

To address any class imbalance in the dataset, the data is resampled to ensure an equal number of samples for each class. The resample method from the pandas library is used for this purpose. Once resampling is complete, the dataset is split into training and testing sets using train_test_split from scikit-learn. The training set is used to train the models, while the testing set is reserved for evaluating the model's performance on unseen data. This step ensures that the model is both trained on a representative sample and tested for generalization on new data.

Step 4: Data Visualization

After the data has been preprocessed and split, the next step is data visualization. Using seaborn, a count plot is generated to visualize the distribution of the target variable. This helps in understanding the class distribution and identifying any imbalances or skewness in the dataset. Visualization plays an important role in identifying potential data issues, such as underrepresented classes, and ensures that the dataset is ready for effective model training.

Step 5: Model Training and Saving

In this step, two machine learning models are trained: Logistic Regression and CatBoost Classifier. The code first checks whether a pre-trained model for each algorithm exists. If a pre-trained model is available, it is loaded from storage. Otherwise, a new model is trained using the training data. The trained model is then saved using joblib for Logistic Regression and CatBoost's built-in functionality for the CatBoost Classifier. Saving the trained model ensures that it can be reused later without the need for retraining, making the process more efficient.

Step 6: Performance Evaluation

Once the models are trained, the next step is to evaluate their performance on the test data. Several performance metrics are calculated, including accuracy, precision, recall, F1-score, and a confusion matrix. These metrics provide a comprehensive assessment of the models' effectiveness in classifying the IoT devices based on the traffic data. The confusion matrix is visualized using seaborn to further understand how well the model distinguishes between the various classes, helping to identify strengths and weaknesses in the classification process.

Step 7: Model Prediction on New Test Data

In the final step, predictions are made on a new test dataset using the trained CatBoost model. The model generates predictions for the classes of IoT devices based on the network traffic metadata provided in the test set. The predicted class labels are then appended to the test data, and the results are printed to display the predicted classes for each entry. This step demonstrates the practical application of the trained machine learning model, showcasing how it can be used for real-time classification and providing insights into the IoT devices generating the network traffic.

3.3 Cat Boost Modelling

CatBoost (Categorical Boosting) is a gradient boosting algorithm specifically designed to handle categorical features and improve model performance by addressing common issues in traditional boosting methods. CatBoost builds an ensemble of decision trees in a sequential manner, where each tree attempts to correct the errors made by the previous trees. In the project, the CatBoost Classifier is used to classify IoT device data. CatBoost handles categorical features natively without the need for

extensive preprocessing or encoding. It employs advanced techniques like ordered boosting and oblivious trees to improve accuracy and reduce overfitting. Ordered boosting helps in reducing the bias introduced by using the same data for training and validation, while oblivious trees ensure that the tree structure is balanced and robust.

CatBoost often exhibits superior performance compared to traditional algorithms like Logistic Regression due to its ability to handle complex relationships and interactions in the data. It effectively manages high-dimensional and heterogeneous datasets, such as those typical in IoT applications, by leveraging its sophisticated gradient boosting techniques. The model's native handling of categorical features and its robust approach to boosting contribute to its enhanced accuracy and ability to generalize well to new data.

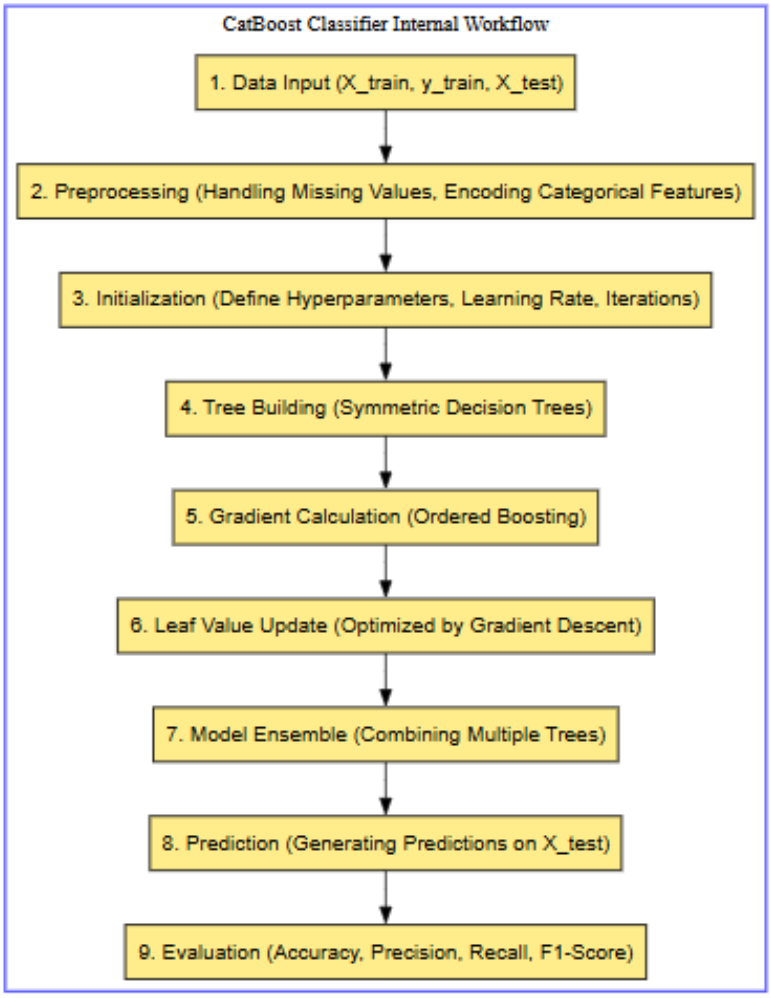


Figure 2: CatBoost classifier internal workflow.

4. RESULTS AND DISCUSSION

The dataset for the IoT data classification project includes several critical columns that help identify and classify IoT devices based on network traffic metadata. The ack column indicates the presence or absence of an acknowledgment signal in communication, while ack_A and ack_B track acknowledgment signals tied to categories 'A' and 'B.' The bytes column measures the size of data packets transmitted, and bytes_A and bytes_B track bytes related to acknowledgment types 'A' and 'B.'

The bytes_A_B_ratio provides a comparative ratio of these byte counts. ds_field_A and ds_field_B describe data fields related to acknowledgment types 'A' and 'B,' adding context to communication patterns, while duration measures the time spent during each communication. URL-related columns such as suffix_is_co.il, suffix_is_com, suffix_is_com.sg, and others track the type of domain or service being accessed, providing insights into the geographic or domain-specific nature of the communication. The suffix_is_else category captures any unspecified URL suffix types, and suffix_is_empty_char_value indicates missing URL suffix data. Additional columns like suffix_is_googleapis.com, suffix_is_net, suffix_is_org, and suffix_is_unresolved track specific URL suffixes and unresolved domain issues, further categorizing traffic. Finally, the device_category column is the target variable, representing the IoT device type, which the machine learning models aim to predict based on the other features. Collectively, these features, ranging from acknowledgment signals to URL suffixes, provide a comprehensive set of data for distinguishing between IoT device types based on network behavior.

Figure 9.5 shows a count plot of the device_category column, illustrating the distribution of different device categories in the dataset. This visualization helps identify any class imbalances that could affect model performance.

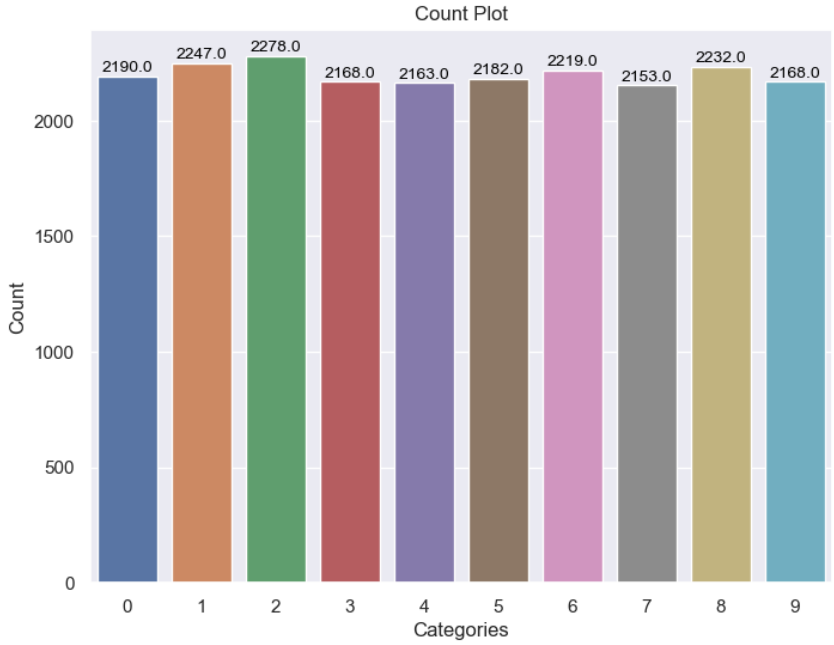
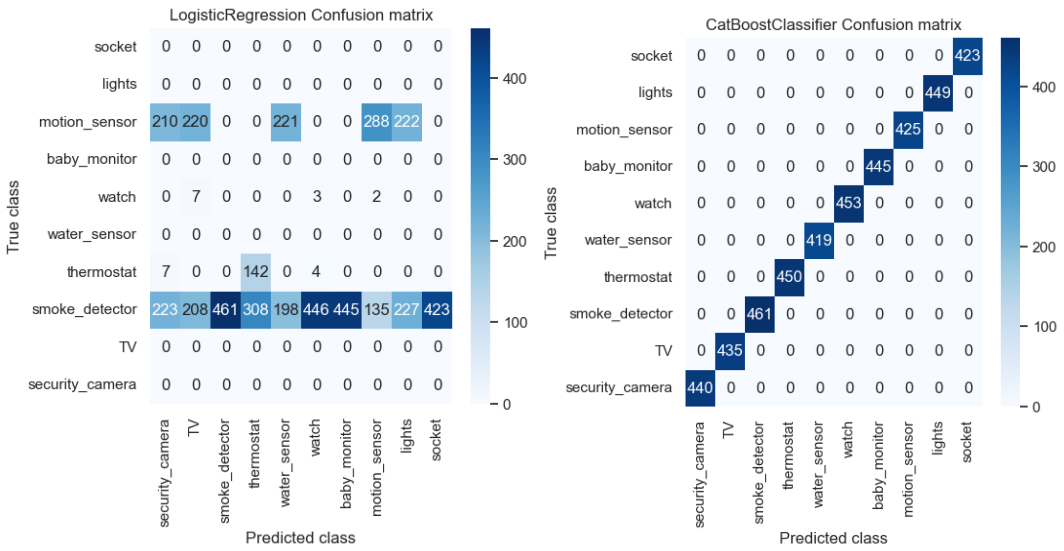


Figure 3: Count Plot of Device Categories



(a) (b)

Fig. 4: Confusion matrices of (a) LRC (b) Cat Boost classifier.

Figure 4 displays the confusion matrix for the Logistic Regression and CatBoost Classifier model. It visualizes the true versus predicted class labels, providing insights into the model's classification performance and highlighting areas where it may struggle and a visual representation of the model's performance in classifying IoT devices, allowing for comparison with the Logistic Regression model and highlighting its superior performance.

Table 1: Performance Comparison for Logistic Regression and CatBoost Classifier

Algorithm Name	Accuracy	Precison	Recall	FScore
Logistic Regression	20.318182	19.998251	15.761341	11.078798
CAT Boost Classifier	100.000000	100.000000	100.000000	100.000000

Table 1 presents a summary of performance metrics including precision, recall, and F1-score for both Logistic Regression and CatBoost Classifier models. This compares the accuracy and effectiveness of the two models, showcasing the superior performance of the CatBoost Classifier.

5. CONCLUSION

The IoT data classification research demonstrates the effectiveness of machine learning models in handling complex, high-dimensional data generated by interconnected devices. Through a structured approach involving data preprocessing, exploratory analysis, model training, and performance evaluation, the research successfully classifies IoT device data into distinct categories. The use of Logistic Regression and CatBoost Classifier models highlights the advancements in machine learning techniques, with the CatBoost Classifier showing superior performance due to its ability to handle categorical features and complex relationships in the data. The comprehensive evaluation metrics, including accuracy, precision, recall, and F1-score, affirm the reliability and robustness of the CatBoost model in providing accurate classifications. The research not only addresses the challenges associated with IoT data classification but also demonstrates the practical application of advanced machine learning methods to real-world problems. The visualizations and performance metrics offer clear

insights into the strengths and limitations of the models, guiding future improvements. Overall, the project underscores the importance of selecting appropriate models and techniques for handling the unique characteristics of IoT data.

REFERENCES

[1] Xie, S.; Zhang, J. Sensor-Based Exercise Rehabilitation Robot Training Method. *J. Sens.* **2023**, *2023*, 7881084.

[2] Qiu, S.; Zhao, H.; Jiang, N.; Wang, Z.; Liu, L.; An, Y.; Zhao, H.; Miao, X.; Liu, R.; Fortino, G. Multi-Sensor Information Fusion Based on Machine Learning for Real Applications in Human Activity Recognition: State-of-the-Art and Research Challenges. *Inf. Fusion* **2022**, *80*, 241–265.

[3] Semwal, V.B.; Gupta, A.; Lalwani, P. An Optimized Hybrid Deep Learning Model Using Ensemble Learning Approach for Human Walking Activities Recognition. *J. Supercomput.* 2021, *77*, 12256–12279.

[4] Prasanth, H.; Caban, M.; Keller, U.; Courtine, G.; Ijspeert, A.; Vallery, H.; Von Zitzewitz, J. Wearable Sensor-Based Real-Time Gait Detection: A Systematic Review. *Sensors* 2021, *21*, 2727.

[5] Yao, S.; Vargas, L.; Hu, X.; Zhu, Y. A Novel Finger Kinematic Tracking Method Based on Skin-Like Wearable Strain Sensors. *IEEE Sens. J.* 2018, *18*, 3010–3015.

[6] Mainali, S.; Darsie, M.E.; Smetana, K.S. Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. *Front. Neurol.* 2021, *12*, 734345.

[7] Mennella, C.; Maniscalco, U.; De Pietro, G.; Esposito, M. The Role of Artificial Intelligence in Future Rehabilitation Services: A Systematic Literature Review. *IEEE Access* 2023.

[8] Liao, Y.; Vakanski, A.; Xian, M.; Paul, D.; Baker, R. A Review of Computational Approaches for Evaluation of Rehabilitation Exercises. *Comput. Biol. Med.* 2020, *119*, 103687.

[9] Wang, Y.; Yang, B.; Hua, Z.; Zhang, J.; Guo, P.; Hao, D.; Gao, Y.; Huang, J. Recent Advancements in Flexible and Wearable Sensors for Biomedical and Healthcare Applications. *J. Phys. D Appl. Phys.* 2022, *55*, 134001.

[10] Cheng, Y.; Wang, K.; Xu, H.; Li, T.; Jin, Q.; Cui, D. Recent Developments in Sensors for Wearable Device Applications. *Anal. Bioanal. Chem.* 2021, *413*, 6037–6057.

[11] Park, Y.-G.; Lee, S.; Park, J.-U. Recent Progress in Wireless Sensors for Wearable Electronics. *Sensors* 2019, *19*, 4353.

[12] Stack, E.; Agarwal, V.; King, R.; Burnett, M.; Tahavori, F.; Janko, B.; Harwin, W.; Ashburn, A.; Kunkel, D. Identifying Balance Impairments in People with Parkinson's Disease Using Video and Wearable Sensors. *Gait Posture* 2018, *62*, 321–326.

[13] Kelly, D.; Esquivel, K.M.; Gillespie, J.; Condell, J.; Davies, R.; Karim, S.; Nevala, E.; Alamäki, A.; Jalovaara, J.; Barton, J.; et al. Feasibility of Sensor Technology for Balance Assessment in Home Rehabilitation Settings. *Sensors* 2021, *21*, 4438

[14] Kimoto, A.; Fujiyama, H.; Machida, M. A Wireless Multi-Layered EMG/MMG/NIRS Sensor for Muscular Activity Evaluation. *Sensors* 2023, *23*, 1539.

[15] Husain, K.; Mohd Zahid, M.S.; Ul Hassan, S.; Hasbullah, S.; Mandala, S. Advances of ECG Sensors from Hardware, Software and Format Interoperability Perspectives. *Electronics* 2021, *10*, 105.