# EL-Based Network Traffic Classification with SDN Clustering for Efficient Operation of Smart Power Grid Enabled Smart Cities

Man Patel*, Naman Jain†, Janam Patel‡, Fenil Ramoliya§, Rajesh Gupta‡, *Member, IEEE*,
Sudeep Tanwar§, *Senior Member, IEEE*,
*†‡§¶‖Department of Computer Science and Engineering, Institute of Technology, Nirma University, Gujarat, 382481, India
Emails: *21bce206@nirmauni.ac.in, †22bce202@nirmauni.ac.in, ‡21bce197@nirmauni.ac.in, §21bce244@nirmauni.ac.in,
¶rajesh.gupta@nirmauni.ac.in, ‖sudeep.tanwar@nirmauni.ac.in,

***Abstract*—Smart power grid enables smart city to be operated on efficient level for sustainable urban planning, economic growth and become an innovation hub. Various types of energy user or provider plants can contribute to providing services while integrated with physical sensor devices, modules, and software interfaces for efficient communication and control. These integrated software interfaces and computer devices also help us during response for any catastrophic condition, fault detection-diagnosis, remote operation-maintenance, integration with Supervisory Control and Data Acquisition (SCADA) systems and provide robust cyber security measures. Software defined network (SDN) allows us to have traffic prioritization, traffic engineering and load balancing, Quality of Service (QoS) enforcement and logical segmentation. However, relying on a single SDN entity to handle all network traffic can introduce several challenges such as single point of failure, scalability limitation, limited redundancy, increased complexity and management overhead and delay in serving of highly critical requests during high-load condition. To mitigate this paramount concern, we propose Ensemble Learning (EL)-based network traffic classification approach with clustering of SDNs to carry-out tasks in efficient and effective manner. The performance evaluation of proposed approach has been done through various performance metrics such as receiver operating characteristic (ROC) curve, precision, recall, f1-score, PR curve and confusion matrix.**

***Index Terms*—Industrial Control System, Smart grids, SDN, EL, Clustering of SDNS**

## I. Introduction

Traditional electric grids are being stretched to their capacity with increasing energy demands and an old fashioned infrastructure. To tackle modern day issues like, cyber threats and environmental impact, that may arise in a smart city, a shift towards a Smart Grid is inevitable. The Smart Grid modifies the preexisting one way interaction between utilities and customers into a two way dialog to exchange not only electricity but also information. It is a developing network of computers, controls, communication and automation aiming to increase availability, reliability and efficiency. With technology enabling sensing along transmission lines and integration of wind and solar energy production, a steady growth can be expected in economic and environmental health.

Software-Defined Networks (SDN), an advanced approach to networking, utilizes controllers and applications to direct traffic in network among a multitude of other features. SDN controllers, the principle elements in any network governed by software, enable centralized flow control for improved network performance. Essentially an operating system for the network, controllers aim to split the control plane and data plane, which in turn allows for automated network management. With the help of APIs, a SDN controller facilitates a two way dialog with network devices, maintaining critical information about entire network, crucial for intelligent decision making. In contrast to traditional networking, SDN controllers reduce inconsistencies and improve flexibility, synchronisation and security in the network.

Current energy industry equipment are not capable to communicate with one and other, there is need for a network structure with numerous IoT devices and software to manage them. Smart Grid, an innovation with numerous advantages, still requires additional security protocols, advanced automation in sensing technology, response to critical disturbances and etc. Utilising a SDN approach to a Smart Grid allows us to store the vast data, collected by the Smart Grid, to a centralized storage for real-time intelligent decision making and rapid response to an error in any device on the grid. SDN integration permits protocol based energy flow control, with automated path selection and operating under heavy flow or scarcity in certain regions. With such extensive real-time data on pricing and other parameters, SDN controllers can communicate with user devices to regulate electricity usage. Incorporating SDN controllers in a Smart Grid further increases load control capacity, like creating individual load profiles through pattern analysis and load shedding ability in times of emergency. Flexibility can be improved through, forecasting and planning on real-time analysis and automated policy changes to avoid physical reconfiguration throughout a power plant or a substation. Tapping into renewable energy sources like, wind and solar energy, will become straightforward using SDN which controls generation, distribution and curtailment.

Further, we discuss research contributions made within recent years for the development in integration of SDN and Smart Grid. Tremendous growth has been marked down towards innovation in network classification and management in a Smart Grid with the assistance of SDN. Zefang Lv *et al.* [1] propose clustering algorithm as well as maintaining privacy

needs within a smart grid ecosystem that generates enormous data. Through the use of both K-means algorithm and K-modes algorithm, with effective privacy budget allocation strategies and relevant loss function, they achieved increased performance and fulfill privacy requirements. Smart grid are susceptible to an increased number of cyber threats causing blackouts and instability. M. Abdelkhalek *et al.* [2] present a solution using Mininet 2.3.0 to create a moving target defense enabled SDN to facilitate communication between centralized control center and Smart Grid. S. Chatzimiltis *et al.* [3] implement a new smart meter intrusion detection system to further tackle security issues posed by conventional substandard centralised practices. Split learning methodology and federated learning approach was analysed and compared to find the best implementation of the intrusion detection system. The use of SDN controllers in Smart Grid has brought upon a multitude of problem including the dilemma related to the placement of the SDN controllers. M. Samir *et al.* [4] use a POCO framework to identify optimum values for all metrics through rigorous simulation, hence discovering best possible controller placement. Different controller placements are generated based on different latency and load modes like, worst node-to-controller latency, worst controller-to-controller latency and controllers load-imbalance. N. A. M. Radzi *et al.* [5] mention the drawback of regular scheduling algorithms using only quality of service (QoS) to prioritise traffic scheduling in a Smart Grid. To adapt to diverse network conditions, they put forward a context aware traffic scheduling (CATSchA) algorithm. It is implemented in a NS-3 simulator with a packet switched network. The algorithm first characterizes the heterogeneous data and then maps it into weighted quality classes, therefore preserving link efficiency , throughput and lowering delay.

As seen in the existing work done my numerous researchers, limited scalability, lack of integration for high-low load dynamic traffic management, issue with quick and reliable response from the system during critical condition can be oberved. To mitigate this pressing concern, we propose Ensemble Learning (EL)-based based network traffic classification approach for efficient energy management in SDN environment using clustering. These clusters are the collection of SDNs to carry out one specific kind of task diverted from ML-implemented mainframe SDN. The performance evaluation includes precision, recall, f1-score comparison for various ML and EL related models and receiver operating characteristic (ROC) curve, precision-recall curve and confusion matrix.

### A. Research Contributions

Following are the research contributions of the proposed framework:

- We proposed a EL-based network classification approach for smart city with Smart Grid to manage network traffic predicaments and improve load management capabilities.
- Division of SDNs into multiple chunks, using clustering based approach, for efficient load management at times of high load and high critical requests from energy-seeking and energy-providing plants.
- The performance of EL-based network classification with clustering of SDN approach is evaluated and simulated with various metrics such as ROC curve, precision, recall, f1-score, PR curve and confusion matrix to show the optimum performance achieved.

### B. Organization of the Paper

The rest of the paper is organized as follows. Section II presents the system model and problem formulation of proposed approach. Section III elaborates the proposed approach in detail with layered architecture. Section IV highlights the performance evaluation. Finally, the paper is concluded with future work in Section V.

## II. System Model and Problem Formulation

This section will provide a detailed explanation of the system model and problem formulation of the proposed framework. The framework is designed for a Smart city, denoted as $C$, which utilizes Smart grids $G$ for efficient energy management in services such as Power plants, Water supply, and traffic management systems. Each Smart grid $G$ is composed of a graph network with multiple nodes $\alpha$, each of which serves a different purpose. Unlike traditional power grids which are centralized systems with unidirectional distribution, the Smart grids $G$ employ advanced communication and control strategies. Additionally, each Smart grid $G$ has various communication links $\beta$ with other nodes, as well as associated traffic $T$.

In the smart city $C$, there exist a network of SDN controllers $S$ through which communication and management of various systems and smart grids are organised. The network consists of multiple SDN nodes $\delta$ according to network topology and services required. Every systems and grids are connected with sensors, and other IoT devices which collects data and transmits with help of $S$ network. The controllers $S$ interacts with each other through communication links $\gamma$ for data transmission.

$$G, S \in C \tag{1}$$

$$G = (\alpha, \beta) \tag{2}$$

$$T \in \alpha \tag{3}$$

$$S = (\delta, \gamma) \tag{4}$$

Our model integrates the Smart grids with SDN, so as to protect against cyber attacks, helps in load balancing, and data packets prioritization. In $S$ consist of a central SDN controller $\Psi_{main}$ which is connected to whole cluster and take networking related decisions. An Ensemble learning based layer $\Lambda_{EL}$ is introduced in the model which is associated with the central controller and classifies the data packets passing through $S$ into $\omega$ classes. Audio files $\omega_{aud}$, Streaming files $\omega_{str}$, Email files $\omega_{mail}$ etc. helps in identifying importance of data packets. Based on traffic through a Smart grid node $T$ and class predicted $\omega$ of data packets, $\Psi_{main}$ prioritise important packets and forward to available node $\eta$ in the SDN cluster.

$$\Psi_{main}, \eta \in S \tag{5}$$

$$\omega = \{\omega_{aud}, \omega_{str}, \omega_{mail}, ...\} \tag{6}$$

We aim to maximize the overall network utility while ensuring that the traffic load on each link does not exceed its capacity. Thus the objective function $\mathcal{J}$ can be repented as in Eq. 7 subject to $R_{ij} \leq C_{ij}, \forall i \in I, \forall j \in J$.

$$\mathcal{J} = \max_{R_{ij}} \sum_{i \in I} \sum_{j \in J} U_{ij}(R_{ij}) \tag{7}$$

here, $R_{ij}$ is the rate of traffic flow from source $i$ to destination $j$ and $C_{ij}$ is the capacity of the link between source $i$ to destination $j$. $U_{ij}(R_{ij})$ represents the utility function for traffic flow from source $i$ to destination $j$. It is a concave function reflecting diminishing marginal utility w.r.t $C_{ij}$.

## III. Proposed Approach

Fig. 1 depicts proposed approach with three main layers: a smart power grid layer, machine learning layer with mainframe SDN and clustered SDN controller layer. (Few images and icon used in 2 were referenced from an online web site [6].)

### A. Smart Power Grid Layer

In an inter-connected world like today's, individual heavily rely smart solutions to increase accuracy, efficiency and productivity in their day-to-day activities. These smart solutions, ranging from the telecommunication industry to the automobile industry, require a constant energy source which is both effectively managed and reliable. In a smart city, there are numerous energy seeking plants and devices, connected in a Smart Grid network, to execute the discussed smart solutions. Solar power plants consist of numerous smaller functioning devices like, temperature sensors photovoltaic panel performance monitors, irradiance sensors and etc. Communication among these devices, supported by the Smart Grid, can be wired or wireless, with data collected by sensors transmitted to a centralized control system for monitoring and decision-making. Using the data and software interfaces, energy production can be tracked, orientation of solar planes can be adjusted as and when required and all other performance can be optimized. With a prominent use of hydroelectric energy in the current world, hydroelectric plants comprise of flow meters, water level sensors, water quality monitors and turbine speed sensors to name a few. Communication takes place through utility-specific protocols pre-defined by the Smart Grid network for seamless coordination between all the devices. Additionally, algorithms are utilised to regulate turbine performance on the basis of water levels, flow rates and other environmental factors, thus providing maximum reliable energy production.

Other than the already discussed energy seeking plants, there exist many other plants such as wind turbine farms, geothermal power plants, charging stations and etc. In a smart city with various energy sources and distribution to numerous energy seeking devices, there needs to be smart supervision to mitigate risk, congestion and failures. Smart Grid provide the required efficient energy management capabilities to a grid with numerous devices through utilisation of all data generated during real-time communication towards predictive modeling and quick recovery from disturbances. With a centralized control system, the Smart Grid allows for decentralized energy generation and automation of grid operations during rapidly changing situations. Allows for introduction of newer technology like, electric vehicle (EV) integration [7], without the hassle of regular physical standard procedure for all devices. Improving load management goes hand in hand with effective energy management, and Smart Grid enables load forecasting and planning to optimize plant operations and manage load independently in an industrial complex using microgrid. Smart Grid empower the grid of energy-seeking devices to adjust energy consumption through the use of grid signals and provide new and improved grid services. Finally, Smart Grid provides innovation in upcoming electric vehicle charging stations, by inspecting metrics like demand, availability and pricing to distribute energy through smart charging stations.

### B. Machine Learning Layer with Mainframe SDN

This section provides information about the dataset used, the preprocessing steps taken, and the classification task. The dataset used for training was VPN-nonVPN (ISCXVPN2016) [8]. It contains real-world network interactions categorized into different classes of network traffic. The dataset consists of a total of 41 features, where 40 are input features for the model. The output $\Theta$ is classified into 20 different classes, such as Gmail and Facebook under the chat class, Vimeo and YouTube under the video class, and so on. The dataset comprises 5988 unique rows, with only 3 null rows that were removed from the dataset. All 20 types of traffic were analyzed and further processed to make them compatible with machine learning models. For better representation and classification, these 20 types were grouped into 7 different types: 'AUDIO', 'FILE', 'VIDEO', 'CHAT', 'STR', 'EMAIL', and 'OTHERS'. After the preprocessing steps, the cleaned data was used for model formulation.

In the network powered by the system model, there is a centralized controller called $\Psi_{main}$. It makes resource allocation decisions based on predefined parameters of the networking. To predict the class of packets, $\Psi_{main}$ uses the Ensemble Learning layer $\Lambda_{EL}$. The data used for training the model was split into training and testing sets to improve learning and evaluation tasks. In order to compare the results, eight different models were evaluated, and a single Ensemble Learning based model was chosen based on its superior performance on evaluation metrics. All models were tested on the test data, and their accuracy, precision, recall, and F1 scores were obtained. This matrix was used to select the best model for the multi-class data.

### C. Clustered SDN Controller Layer

$\Psi_{main}$ operates as the central entity to divert network traffic through $\Theta$ predicted from $\Lambda_{EL}$. $\Psi_{main}$ is used to perform load balancing in the network to serve highly critical request coming from industrial infrastructure. As any power plant, smart transportation hubs, telecommunication infrastructure, waste-to-energy facilities, etc., require communication and control only through software based devices. The other traffic

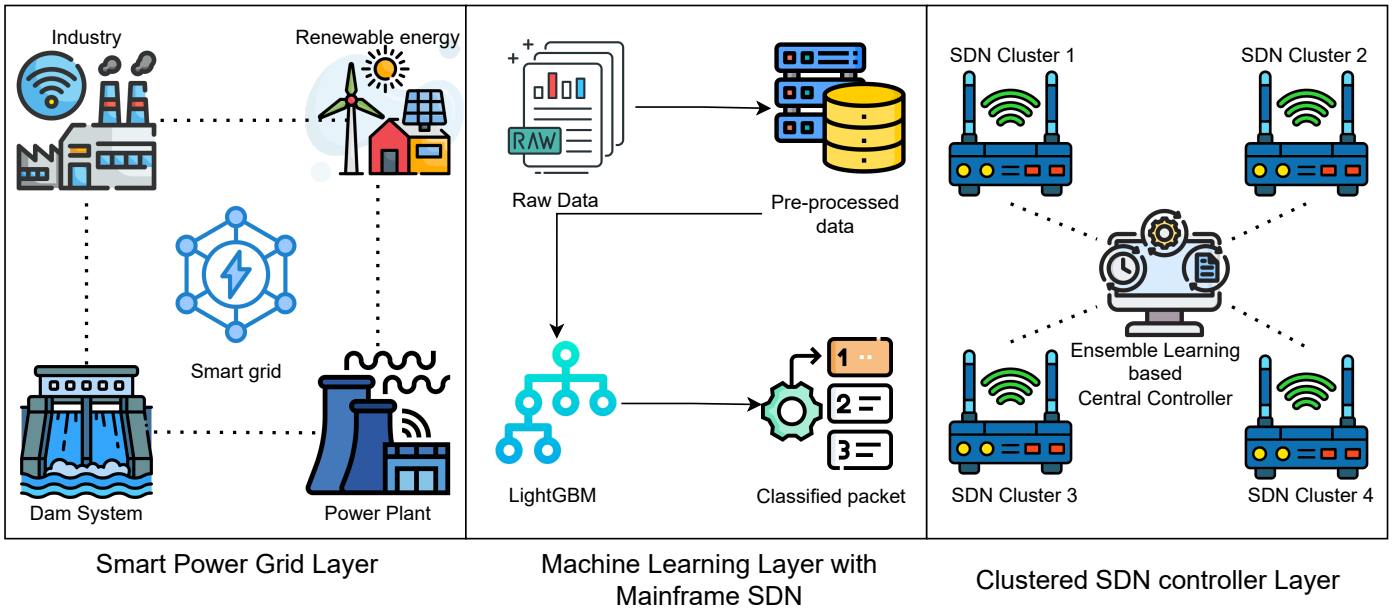



Fig. 1: Proposed Approach Architecture

generated from these working devices when they are not serving to this smart grid energy-seeking plants, needs to be ready to handle any critical incoming data. Here, our proposed approach comes in to picture which demonstrates clustering of SDNs to perform one specific type of data handling job, which is much more preferable over only one SDN serving, processing and delivering requests in this smart city enabled scenario.

Using routing algorithm $\Psi_{main}$ redirect the classified traffic to the corresponding SDN cluster ($\Psi_k^i$, where $k$ is class of cluster with $i$ number of SDNs) specialized in handling that particular type of data. One possible edge case in the given proposed approach could be the 'low load' in the network. For this purpose, the SDN cluster encompasses an additional type of smaller SDN containing cluster referred to as the default cluster ($\Psi_{def}$). Which is able to handle any kind of request coming from network for processing and passing when overall load in the network is low. $\Psi_{def}$ consist of SDNs which are capable of doing all jobs but in limited number of instances. During high load condition, clustering based traffic divergence mechanism using $\Lambda_{EL}$'s prediction perform the load balancing task. For that purpose cluster's can accommodate queuing theory of load balancing with $\lambda_k$ as the arrival rate of requests to the $p^{th}$ SDN cluster and $\mu_k$ as the service rate of the $p^{th}$ SDN cluster. As per M/M/1 queue model, the traffic intensity of the $p^{th}$ cluster can be defined as in Eq. 8

$$\rho_k^p = \frac{\lambda_k^p}{\mu_k^p} \tag{8}$$

For $\Lambda_{EL}$ and $\Psi_k^i$ in smart city scenario with mix of regular traffic requests and critical energy plant requests it is require to have probability of the system being empty and probability of having $m$ requests in the system to carry out effective load balancing. Eq. 9 represent probability $\mathcal{P}_0$ of system being empty with no request incoming or very low number of request incoming. Eq. 10 showcase the probability $\mathcal{P}_{NN}$ of the system with more than $N$ requests and Eq. 11 showcase the probability $\mathcal{P}_{nn}$ of the system with less than $N$ requests for their cluster.

$$\mathcal{P}_0 = \frac{1}{1 + \sum_{n=1}^{N-1} \frac{(N\rho^k)^n}{n!} + \frac{(N\rho^k)^N}{N!(1-\rho^k)}} \tag{9}$$

$$\mathcal{P}_{NN} = \frac{N^N \rho^k}{N!(N-\rho^k)^n} \mathcal{P}_0 \tag{10}$$

$$\mathcal{P}_{nn} = \frac{(N\rho^k)^n}{n!} \mathcal{P}_0 \tag{11}$$

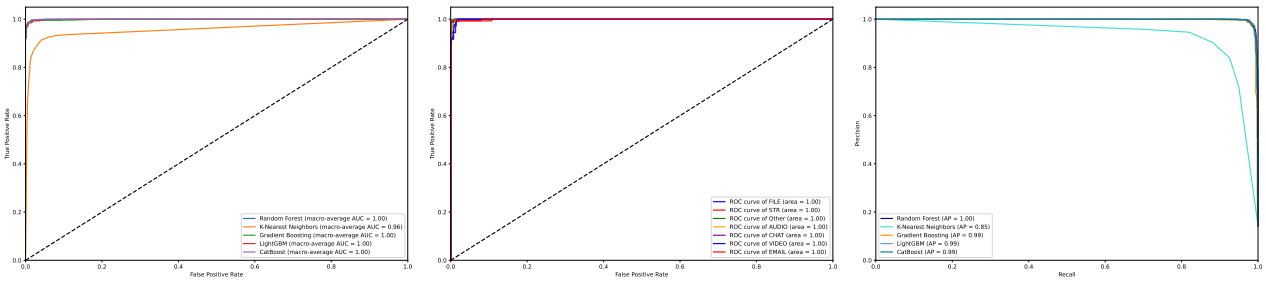
## IV. Performance Evaluation

### A. Simulation setup and tools

The system model has been simulated on Jupyter Notebook using Python v3.9.6 for the entire approach. Pandas v2.0.2 is used to carry out data pre-processing. The operations on arrays are performed using Numpy v1.23.5, while the data analysis and results visualization are carried out using Matplotlib v3.7.1. The system used for simulation is Apple Mac M1, which has an 8-core CPU, 8-core GPU, and 8GB RAM. Using consistent configurations increases reliability and trustworthiness of evaluation and testing. The proposed model and other ML models referenced in table I were trained on the default hyperparameters for all the algorithms.

### B. Performance analysis

In this section, we present a comprehensive analysis of the system model, comparing it with other machine learning algorithms. The table I shows the performance metrics of our




(a) Macro averaged ROC. (b) ROC-AUC curve. (c) PR curve.

Fig. 2: (a) Macro-averaged ROC-AUC comparison for models, (b) ROC curve for classes predicted (c) PR comparison curve for models.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **ROC-AUC** |
|---|---|---|---|---|---|
| Logistic Regression | 0.287 | 0.082 | 0.287 | 0.128 | 0.500 |
| Decision Tree | 0.974 | 0.974 | 0.974 | 0.974 | 0.985 |
| Random Forest | 0.979 | 0.980 | 0.979 | 0.979 | 0.9995 |
| K-Nearest Neighbors | 0.895 | 0.895 | 0.895 | 0.894 | 0.970 |
| Multi. Naive Bayes | 0.226 | 0.304 | 0.226 | 0.191 | 0.551 |
| Gradient Boosting | 0.979 | 0.980 | 0.979 | 0.979 | 0.9996 |
| CatBoost | 0.982 | 0.983 | 0.982 | 0.982 | 0.9996 |
| **LightGBM** | **0.983** | **0.983** | **0.983** | **0.983** | **0.9997** |

TABLE I: Performance Metrics of Various Models

software-defined networking (SDN) classification. The classification task achieved 98.3% accuracy, with LightGBM outperforming other models. Apart from accuracy, other evaluation metrics such as precision, recall, F1-score, and ROC-AUC are also promising. This shows the effectiveness of our proposed model compared to other models. The LightGBM ensemble method outperformed all other algorithms in accurately predicting the classes. This is due to its strong predictive ability. Additionally, the leaf-wise growth of the algorithm resulted in a more complex tree structure that can capture finer patterns in the data.

Fig. 2a and Fig. 2b represents the Receiver Operating Characteristic (ROC) curve for our a multiclass SDN classification problem. The ROC curve evaluates the performance of the classifier across different thresholds for classifying instances into multiple classes. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for various threshold values. TPR is also known as sensitivity or recall, while FPR is the complement of specificity. The Area Under the Curve (AUC) measures the overall performance of the model. The higher the AUC value, the better the classifier's ability to distinguish between multiple classes.The Fig. 2b displays the ROC curve for seven classes, which indicates an outstanding performance. The Fig. 2a also shows the macro-average AUC for the top five algorithms that were used to evaluate the performance over multiclass classifiers. This metric helps to assess the model's performance in cases where the classes are imbalanced or have equal importance. It shows the classifier's generalization across all classes without any bias towards any specific class. Fig. 3 shows the confusion matrix, which provides a better understanding of how the system model is performing on the test data for each class.

Due to the class imbalance, Precision-Recall (PR) curve was additionally used for performance evaluation. Precision and recall are two important metrics used in evaluating the performance of prediction models. Precision measures the percentage of correctly predicted positive instances among all instances that are predicted as positive, while recall measures the percentage of correctly predicted positive instances among all actual positive instances. These metrics are particularly significant when the cost of false positives and false negatives differs. The Fig. 2c displays the Precision-Recall (PR) curve for multiclass classifiers. To evaluate the performance of the model across different classes, the PR curve is calculated for each class individually. Similar to the macro-average AUC in ROC analysis, the code calculates a macro-average PR curve, where precision and recall are averaged across all classes. This macro-average PR curve provides an overall evaluation of the model's performance across all classes, treating each class equally. The area under the PR curve is computed for each class and then averaged to obtain a single numerical value that summarizes the model's performance. It helps in selecting an appropriate threshold that balances precision and recall based on the specific requirements of the application.

## V. Conclusion

In this paper, we propose clustering based approach to have similar single job serving SDNs in one single cluster. The mainframe SDN capable of EL capacity for network traffic classification act as central entity which diverts the traffic to its appropriate cluster. In the smart city, where highly sustainable yet capable plant working with large amount of energy often require communication and control of sensors, physical modules and devices through software interface or computer entity.

These devices are also connected to the outside work network and perform their routine job of monitoring, analyzing, and responding to environmental conditions, operational parameters, and user inputs in real-time. Our proposed approach helps the overall ecosystem by efficiently classifying and processing incoming-outgoing network data while serving at peak during critical conditions and high propriety request from energy plant. Future scope involves integration of edge computing, enhanced security and privacy measures and federated learning-based dynamic resource allocation strategies.

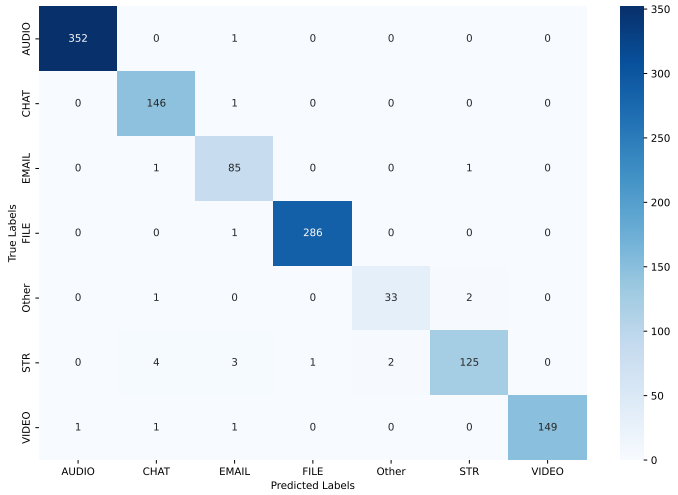



Fig. 3: Confusion Matrix for predicted data class using LightBGM

## REFERENCES


[1] Z. Lv, L. Wang, Z. Guan, J. Wu, X. Du, H. Zhao, and M. Guizani, "An optimizing and differentially private clustering algorithm for mixed data in sdn-based smart grid," *IEEE Access*, vol. 7, pp. 45773–45782, 2019.

[2] M. Abdelkhalek, B. Hyder, M. Govindarasu, and C. G. Rieger, "Moving target defense routing for sdn-enabled smart grid," in *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 215–220, 2022.

[3] S. Chatzimiltis, M. Shojafar, M. B. Mashhadi, and R. Tafazolli, "A collaborative software defined network-based smart grid intrusion detection system," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 700–711, 2024.

[4] M. Samir, E. Samir, M. Azab, M. R. M. Rizk, and N. Sadek, "Towards optimal placement of controllers in sdn-enabled smart grid," in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 95–100, 2018.

[5] N. A. M. Radzi, N. Suhaimy, W. S. H. M. W. Ahmad, A. Ismail, F. Abdullah, M. Z. Jamaludin, and M. N. Zakaria, "Context aware traffic scheduling algorithm for power distribution smart grid network," *IEEE Access*, vol. 7, pp. 104072–104084, 2019.

[6] "Flaticon." https://www.flaticon.com/, 2023. Accessed on September 12, 2023.

[7] H. Patil and V. N. Kalkhambkar, "Grid integration of electric vehicles for economic benefits: A review," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 1, pp. 13–26, 2021.

[8] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and vpn traffic using time-related," in *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, pp. 407–414, 2016.