Edge Mining on IoT Devices Using Anomaly Detection

Kavin Kamaraj*, Behnam Dezfouli†, and Yuhong Liu‡
Internet of Things Research Lab, Department of Computer Science and Engineering, Santa Clara University, USA
*kkamaraj@scu.edu, †bdezfouli@scu.edu, ‡yliu@scu.edu,

***Abstract*—With continuous monitoring and sensing, millions of Internet of Things sensors all over the world generate tremendous amounts of data every minute. As a result, recent studies start to raise the question as whether to send all the sensing data directly to the cloud (i.e., direct transmission), or to preprocess such data at the network edge and only send necessary data to the cloud (i.e., preprocessing at the edge). Anomaly detection is particularly useful as an edge mining technique to reduce the transmission overhead in such a context when the frequently monitored activities contain only a sparse set of anomalies. This paper analyzes the potential overhead-savings of machine learning based anomaly detection models on the edge in three different IoT scenarios. Our experimental results prove that by choosing the appropriate anomaly detection models, we are able to effectively reduce the total amount of transmission energy as well as minimize required cloud storage. We prove that Random Forest, Multilayer Perceptron, and Discriminant Analysis models can viably save time and energy on the edge device during data transmission. K-Nearest Neighbors, although reliable in terms of prediction accuracy, demands exorbitant overhead and results in net time and energy loss on the edge device. In addition to presenting our model results for the different IoT scenarios, we provide guidelines for potential model selections through analysis of involved tradeoffs such as training overhead, prediction overhead, and classification accuracy.**

***Keywords*—Edge mining, fog computing, anomaly detection, supervised machine learning**

I. Introduction

The Internet of Things (IoT) refers to a network of billions of interconnected devices that have the ability to communicate and exchange data over the Internet. Such IoT devices range from sensors, smart phones, computers, vehicles, building appliances, and health devices. The extension of Internet connectivity to physical devices and everyday objects has substantially increased worldwide real-time data collection and transmission. As of 2019 there are approximately 9 billion IoT devices across the world and by 2020 this number will surge to over 25 billion [1]. These IoT edge devices, such as smart light bulbs, wearable medical devices, doors, heaters, sensors for smart agriculture, etc., which typically host a variety of sensors for temperature, pressure, humidity, light, motion and acceleration, are often resource-constrained[2]. For example, many of them are battery-powered, with limited processing power which is just sufficient for the task at hand so that they can be mass-produced while minimizing costs.

On the other hand, the amount of data produced by these sensors at this scale is staggering. As IoT devices grow in number, the tremendous amount of sensing data collected has raised great challenges for data transmission overhead (time and energy) and cloud storage. Applications of cost-cutting edge mining techniques to reduce packet transmission and remote storage requirements are rapidly growing throughout IoT networks [3], [4]. One of the most basic edge mining techniques is random sampling to reduce the number of observations sent to the cloud. More sophisticated methods, such as anomaly detection, isolate and transmit only the contextually relevant observations [5]. The guiding principle behind anomaly detection is that only unexpected behavior needs to be notified to the centralized cloud. Contemporary works specify different anomaly detection methods ranging from basic thresholding to machine learning algorithms [6], [7]. However, as the resources available at each IoT edge device can be rather limited in terms of power, memory, connectivity, bandwidth, and computation, it is critical to choose appropriate anomaly detection algorithms that can not only effectively identify abnormal behaviors but also consume limited resources at the edge devices.

In this work we present supervised machine learning based anomaly detection that can substantially reduce energy consumption and transmission overhead on the edge and storage requirements on the cloud. We conduct our experiments on a custom testbed inclusive of an edge device, the cloud, and an energy measurement platform. Four classes of machine learning algorithms, Random Forest Classifier (RF), Multilayer Perceptron Classifier (MLP), K-Nearest Neighbor Classifier (KNN), and Discriminant Analysis Classifier (DA) are benchmarked on a Raspberry Pi 3 (RPi3) edge device for different anomaly detection scenarios. We use both Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) variants of DA for this study. Our work makes several novel contributions to anomaly detection used in the context of edge mining. First, we benchmark training and prediction phase overhead (i.e., time and energy consumption) at the edge device for each model on multiple datasets. Second, real data rather than synthetic data have been adopted for experiments. Third, we demonstrate tangible transmission cost-savings using machine learning based anomaly detection for multiple datasets.

We demonstrate significant overhead-savings achieved using MLP, RF, and DA anomaly detection model classes during the data processing and transmission period. Furthermore, we identify the best anomaly detection model for different application scenarios. For example, MLP features one of the shortest prediction phases, which is recommended for time-

sensitive scenarios. However, it also has one of the most costly training phases which is undesirable if there are resource constraints for training. By analyzing these tradeoffs, we better understand why the models achieve different levels of performance in different scenarios.

The rest of the paper is organized as follows. In Section II, we present related work focusing on previously studied applications of anomaly detection on IoT edge devices. Section III discusses the supervised models and the datasets used in our experiments. Section IV contains both experimental procedure and results. Finally, Section V concludes the paper and provides future research directions.

II. Related Work

Anomaly detection is one of the the most popular edge mining techniques explored in IoT scenarios [8], [9]. In [10], [11], [12], and [13] anomaly detection is used as a method to implement an Intrusion Detection System (IDS) for Wireless Sensor Networks (WSN). Sommer et al. [12] propose the use of LDA to reduce the dimensionality of network intrusion datasets and applies both Naive Bayes and KNN algorithms for anomaly classification. In [13], the authors benchmark the performance of anomaly detection (i.e., false positive rates) using an unsupervised outlier detection technique based on the RF algorithm. Furthermore, [14] demonstrates the effectiveness of autoencoders for an unsupervised IDS and proposes a novel splitting and learning mechanism to lower false positive detection. Although these works explore novel applications of anomaly detection on the IoT edge, they do not focus on resource constrained scenarios and therefore do not delve into the time and energy consumption of these methods.

In [15], [16], and [17] anomaly detection is investigated in healthcare applications. Arijit et. al [16] propose cardiac anomaly detection with low false negative counts and stress the importance of capturing outliers in healthcare applications. Similar to the IDS studies, this work focuses extensively on anomaly detection implementation but does not consider the factor of resource consumption. Several works also tackle anomaly detection in IoT applications outside of IDS and healthcare. In [18] non-machine learning anomaly detection algorithms are proposed for a set of heterogenous sensors in an IoT WSN. In [19] an autoencoder neural network is used for determining anomaly readings from a testbed of eight temperature and humidity sensors. These works also, however, do not consider overhead nor consider pros and cons using different anomaly detection methods.

Few works consider resource constrained IoT scenarios. For example, Sedjelmaci et. al [20] test a reputation model based on game theory to predict attack signatures on a resource constrained IoT device. In addition to this, the work of Lyu et. al [21] is one of the few works which presents cost-savings potential of anomaly detection using both real and synthetic datasets on a resource constrained IoT platform. However, this work only evaluates unsupervised hyperellipsoidal clustering and does not include supervised machine learning algorithms. Unsupervised methods are useful in the absence of ground-truth information but are generally not as accurate as supervised methods for classification tasks.

TABLE I: Mathematical notations and symbols.

d	Number of features
h	Number of neurons per hidden layer
i	Number of iterations
k	Number of hidden layers
K	Number of neighbors
M	Number of testing samples
N	Number of training samples
o	Number of output neurons
P	Distance-metric complexity
T	Number of trees

TABLE II: Training and prediction time complexity of anomaly detection models.

Classifier	Training Complexity	Testing Complexity
RF	$O(N * \sqrt{(d * T)})$	$O(Td)$
MLP	$O(Ndh^k oi)$	$O(Nhk)$
KNN	$O(1)$	$O(NPMlogK)$
LDA	if $N > d$: $O(Nd^2)$ if $d > N$: $O(d^3)$	$O(M)$
QDA	$O(d^4)$	$O(M \log M)$

III. Anomaly Detection Models and Datasets Used

In this section below, we provide an analysis of each machine learning method benchmarked in our study and discuss their general use cases, time complexity, and performance tradeoffs involved in their training and prediction phases. We have chosen RF, MLP, KNN, LDA, and QDA (i.e., two variants of DA) for this study because they are among the most popular machine learning classification methods [22]. Furthermore, all chosen models have distinct underlying mathematical mechanisms which account for varying model performance across different scenarios. Table I denotes all mathematical notations. Table II provides each model's *theoretical* training and testing time complexity.

A. Random Forests

RF models are applicable to a wide range of classification problems [22]. In a random forest, each node is split using a subset of features randomly chosen at that node. This strategy is robust against overfitting and enables RF to perform better than many other classifiers, including discriminant analysis, support vector machines, and neural networks [22]. The only major downside of RF is that a large number of trees can slow down the algorithm for real-time predictions. Suppose there are T randomized trees, d features, and N training samples, RF's training time complexity is $O(N^2\sqrt{dT})$ [23]. The computational complexity at test time for a RF with T trees and d features is $O(Td)$ [23].

B. Multilayer Perceptron

MLP is the most known and frequently used type of neural network using the backpropagation training algorithm. In recent years, neural networks have been extensively used for pattern recognition and optimization. MLP models contain three types of layers: input layer, output layer, and hidden layer. Each node in the input layer, from top to bottom, passes an input data point to each neuron in the first hidden layer. Then, each hidden layer neuron multiplies each value with a

TABLE III: Overview of the datasets.

Dataset	Dimensionality	% of Anomalies	Anomaly Type	IoT Scenario
KDDCup	567497x3	0.35%	Attack on the Network	Intrusion Detection
Digits	30000x784	11%	Digit 0	Anomalous Image Detection
Gestures	11674x64	25.3%	A Hand Contraction of Patient	Health Monitoring

weight vector and computes the sum of the multiplied values. Subsequently each hidden layer neuron applies its activation function to this sum, and sends the resulting value to the next layer and eventually to the output layer. Suppose there are N training samples, d features, k hidden layers each containing h neurons, o output neurons, and i iterations. The time complexity of MLP training is $O(Ndh^koi)$ [24]. It is advisable to start with a small number of hidden layers and nodes given the high time complexity of MLP's backpropagation algorithm and time-consuming grid search procedure [24]. One of the most notable advantages of MLP, is its low prediction complexity, $O(Nhk)$, which makes it suitable for time-sensitive anomaly detection tasks.

C. K-Nearest Neighbors

K-Nearest Neighbor Classifier (KNN) is a relatively simple learning algorithm. It is very commonly used in text mining, agricultural predictions, and stock market forecasting [25]. Because KNN does not make any assumptions about the underlying data distribution, it is particularly suitable for applications with little or no prior knowledge about the distribution of the dataset.

KNNs are generally reputed for high prediction accuracy with respect to precision and recall. Furthermore, the training phase of KNN classifiers is very efficient. The training time complexity of KNN is only $O(1)$. Nevertheless, there are several drawbacks of KNNs. First, the model has high space complexity as it stores all training data instances. Second, finding the most optimal value for K is not trivial [26]. KNN's most significant drawback, however, is its exorbitant prediction phase overhead, which is $O(NPMlogK)$ with N training samples, M test samples, K neighbors, and P distance metric time complexity. The high overhead at the prediction phase may make it less suitable to be carried on by resource constrained IoT edge devices.

D. Discriminant Analysis

In this work, we use both LDA and QDA. LDA first projects a dataset onto lower-dimensional space to prevent overfitting and to generate linear class-separability. QDA also performs dimensionality reduction but generates a quadratic line to fit the training data. [27]. Despite its potential for non-linear data patterns, the number of the parameters needed by QDA scales quadratically with that of the variables, making it slower for very high dimensional datasets. Both LDA and QDA are extensively used for bankruptcy status classification and face classification. LDA's training complexity is $O(d^3)$ if $d > N$ (i.e., more features than training samples) and $O(Nd^2)$ if $N < d$ (i.e., more training samples than features). QDA's training complexity is generally $O(d^4)$ [27]. LDA's prediction complexity is $O(N)$ whereas QDA's prediction complexity is $O(N^2)$.

TABLE IV: Training and prediction data dimensions.

Dataset	Training Data Size	Prediction Buffer Size
KDDCup	10000	50000
Digits	3000	2000
Gestures	3000	3000

E. Datasets Used

In this paper, we benchmark all four classes of classification algorithms on the following datasets: KDDCup 1999 (KDDCup) [28], Digits 0-9 (Digits) [29], and Hand Gestures (Gestures) [30]. We have chosen these datasets for the following reasons. First, they represent different application scenarios. Second, they represent different percentages of anomalies at 0.35%, 10%, and 25.3% respectively. Third, they represent different orders of dimensionality at 3, 784, and 64 respectively. Table III provides an overview of each dataset with regard to anomaly detection scenario, percentage of anomalies, and dimensionality. The original KDDCup dataset from UCI machine learning repository contains 41 attributes. ODDS Library from Stonybrook University, New York has reduced the dataset to 3 attributes: duration of communication, number of incoming bytes, and number of outgoing bytes. Each training sample contains 3 features and an output value indicating whether the network is in a secure or compromised state. The original data set has 3,925,651 attacks (80.1%) out of 4,898,431 records. ODDS has forged this dataset to contain only 3,377 attacks (0.35%) out of 567,497 records. The end goal in this scenario is to only notify the cloud of the sparsely occurring attacks on the network.

Digits is a dataset containing images (28×28) of digits between 0-9. There are a total of 784 features with each feature corresponding to each constituent pixel of a given image. Digit 0 is considered the anomalous class and constitutes roughly 10% of the dataset. We consider digit 0 to represent an unexpected entity or intruder captured within a collection of image frames in a video stream. Anomaly detection can substantially lower transmission overhead when filtration is applied to high dimensional data such as images.

Hand Gestures is a dataset with 64 total attributes each representing sensor measurements of a person's hand while playing the game of rock-paper-scissors. The original dataset classifies each observed motion as either rock (closed fist), paper (open fist), scissors (two fingers pointed), and neutral (flat hand). For our study, we designate rock as the anomalous gesture and group the other three gestures into the norm class. We consider the rock symbol to be emblematic of a patient's hand contraction which requires medical assistance. In this process, we introduce a scenario where 25.3% of the closed fist instances (i.e., anomalies) are representative of a health condition requiring notification to the cloud.

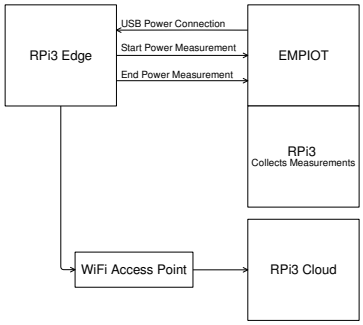

Fig. 1: The testbed used for our experiment inclusive of three RPi3s and the EMPIOT energy measurement platform. The RPi3 hosting EMPIOT captures energy measurements of the edge RPi3 transmitting data to the cloud RPi3.

IV. Experimentation and Analysis

The organization of this section is as follows. First, we specify our testbed, model creation process, and model benchmarking process. Then, we show how each anomaly detection model performs on each dataset (i.e., scenario) with respect to both overhead and anomaly detection accuracy. Furthermore, we show the best model for each scenario and analyze tradeoffs involved in model selection such as training overhead, prediction overhead, and prediction accuracy.

A. Methodology

The testbed includes three RPi3s, which are used as the edge device, the cloud, and the interface to connect to our energy measurement platform, EMPIOT [31]. All RPi3s feature a BCM2837 SoC, a 1.2 GHZ quad-core ARM Cortex A53 processor, and 1 GB LPDDR2-900 SDRAM and run Debian OS [32]. Both the client and cloud RPi3s connect to an access point (i.e., router) through an 802.11 link. We run and benchmark all machine learning algorithms for anomaly detection on the edge RPi3. Time measurements are captured by setting timestamps in the Python code before and after algorithm execution. Energy measurements are captured by EMPIOT. EMPIOT is composed of a shield and is installed on top of the host RPi3. Figure 1 shows our experimental setup inclusive of edge, cloud, and EMPIOT.

We use scikit-learn implementations of RF, MLP, KNN, LDA, and QDA for our anomaly detection experiment. There is minimal data preprocessing for the three datasets used in this experiment because they do not contain missing column values nor incorrect formatting. All code related to anomaly detection such as data aggregation and model training/validation is written in Python 3.6 making use of bcrypt, pandas, numpy, and scikit-learn libraries. For each dataset, we first tune and validate each model by running an offline grid search (i.e., hyperparameter tuning). We configure the grid search to rank model configurations based on precision and recall. Note that we consider grid search to be a purely offline operation performed on the cloud and do not consider its overhead. We also note that grid search is not necessarily exhaustive in all scenarios because not all models require extensive hyperparameter tuning. For example, RF generally demonstrates high prediction accuracy on all datasets using default scikit-learn parameters. MLP, on the other hand, requires extensive hyperparameter tuning to converge on hidden layer and node count. Having decided on hyperparameter configuration, we are left with one-time model training. For each scenario, we train a model with the least number of samples that enables it to predict with both precision and recall greater than 80% and either precision or recall greater than 90%. Table IV shows the training data size for each dataset. We choose to benchmark model training on the RPi3 edge to provide reference for cases when the edge needs to carry on model training. Note that we also consider model training to be an offline operation and therefore will not factor this cost into our overhead-savings analysis.

We benchmark model precision and recall for each dataset. Precision indicates the percentage of relevant results $Precision = \frac{TruePositiveCount}{TruePositiveCount+FalsePositiveCount}$. Recall indicates the percentage of results accurately classified by the algorithm $Recall = \frac{TruePositiveCount}{TruePositiveCount+FalseNegativeCount}$. We also tabulate the time and energy consumption of each model's training phase and prediction phase. The training phase entails instantiating a model with a hyperparameter configuration and fitting the model on training samples. The prediction phase entails executing the model predict function on a buffer of test samples. Time is measured by setting timers within the Python code while energy is measured using EMPIOT. All data transmitted to the cloud is secured using 256-bit AES encryption. Note that encryption overhead is factored into our cost analysis. For each model and dataset, we show precision, recall, prediction time and energy, training time and energy, the number of observations and MBs saved by the cloud, and lastly, the time and energy saved using anomaly detection. We then closely analyze the quantitative results and study the performance tradeoffs of different models in different scenarios.

B. Results and Analysis

We present the model performance and model overhead results obtained from each dataset. We recommend the best model for each scenario based on training and prediction overhead and model accuracy. Note that the time and energy savings reported for each anomaly detection model consider prediction cost and not the offline training cost. We conclude this section with an analysis of general model behavior observed across all datasets and propose recommended use cases for each model.

1) KDDCup Dataset: For the 3 feature KDDCup dataset, all models are trained using 10000 network observations out of a total of 567497 labeled training samples with 0.35% anomaly rate. Note that this dataset has the lowest dimensionality as well as the lowest anomaly rate among the three datasets. In this scenario, the edge device aggregates 50000 size network status buffers for the prediction phase (i.e., anomaly detection) and subsequent transmission.

Figure 2 shows the precision and recall values observed when applying each model for each scenario. We see that all models post exceptional precision and recall values on the

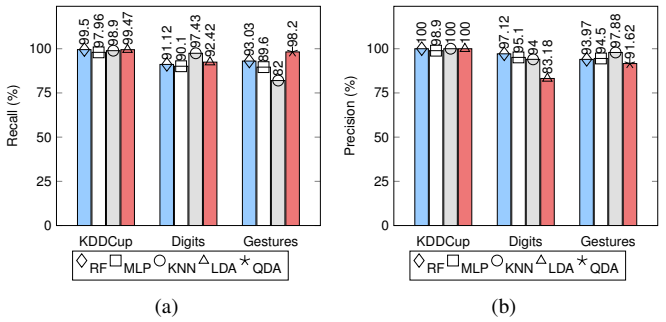

Fig. 2: All models are viable for the datasets chosen. All models perform best for the KDDCup dataset, containing the lowest number of features, and post precision and recall values of 98.9% and 97.96% respectively. RF and QDA offer the best prediction accuracy for Gestures while RF and KNN offer the best prediction accuracy for Digits.

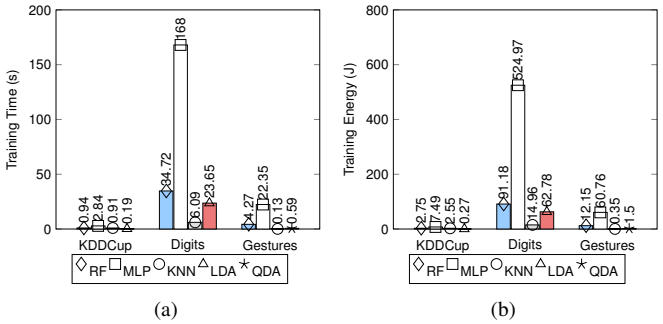

Fig. 3: MLP features the most expensive training phase with RF, QDA, LDA, and KNN generally following in this order as shown in (a) and (b). MLP costs approximately 500% more time and 500% more energy than the next best performing model across all datasets.

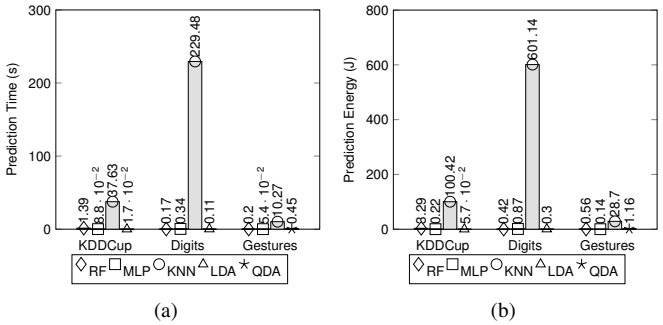

Fig. 4: LDA offers the most cost-effective prediction phase for both Digits as well as KDDCup datasets as shown in (a) and (b). MLP offers the most cost-effective prediction phase for the 64 feature Gestures dataset. KNN costs nearly 600% more energy and 650% more time than the next best performing model across all datasets.

sparse anomaly KDDCup dataset. Relatively speaking, MLP posts the lowest precision and recall values at 98.9% and 97.9% respectively. All other models post over 99% recall and 100% precision. We subsequently examined that one of the features in this dataset has a Gaussian distribution and that the anomalies primarily occur when the feature's value lied $\pm 3\sigma$ outside of the mean. Given this level of accuracy in anomaly detection, we assert that the edge device can save around 99.5% of the data (i.e., approximately 1.2 MB) per buffer sent to the cloud.

Figure 3 shows the training time and energy for each model applied on different datasets. Specifically for KDDCup, LDA has the most time and energy efficient training phase among all models. This is because LDA uses an underlying dimensionality-reduction technique for generating a class-separating boundary that is efficiently performed on a 3 dimensional dataset. KNN, RF, and MLP follow in order. RF takes a significantly longer time proportional to KNN for this dataset due to the relatively large test data buffer size.

Figure 4 shows the prediction time and energy when each model is applied. As shown in Figure 4, when applied on the KDDCup data, LDA is also the most cost-effective model for prediction. The low dimensionality of inputs to the model enables LDA to classify anomalies very efficiently. It is followed by MLP, RF, and KNN. MLP is marginally worse than LDA for this scenario considering the fact that it is slightly inferior in terms of prediction accuracy. RF is a well-rounded choice with respect to both prediction accuracy and overhead. Note that RF's prediction overhead exceeds LDA's prediction overhead due to the complexity involved in processing test samples at multiple nodes at multiple tree levels. KNN is rendered slow and ineffective for this scenario as its prediction phase consumes nearly 38 seconds and 101J for a 50000×3 buffer.

2) Digits Dataset: The 784 feature Digits dataset has the highest dimensionality among all our datasets and is emblematic of an anomalous image detection scenario. For this dataset, all models are trained using 3000 images out of the total 20000 images with a 10% anomaly rate. The edge device performs anomaly detection on buffers containing 2000 images and transmits the anomalous images identified.

Regarding prediction accuracy, Figure 2 shows that KNN offers the best overall precision and recall among all models at 94% and 97.4% respectively. RF posts the highest precision out of all the models at 97.1%. LDA performs the poorest in this scenario offering 92.42% recall but only 83.18% precision. Given the overall high level of accuracy in anomaly detection, we assert that the edge device can save around 89.5% of the data (i.e., approximately 11.3 MB) per buffer sent to the cloud.

As far as training overhead, Figure 3 shows that KNN has the most efficient training phase on the Digits dataset. KNN only stores training samples as part of its training phase rather than formulating a mapping between inputs and outputs. Therefore, for this 784 dimensional dataset, the other algorithms have a considerably more demanding training phase. KNN is followed by LDA, RF, and MLP in order. MLP, which has the most expensive training phase, costs 168 seconds and 524.97J. This result is expected given the computationally expensive nature of MLP's backpropagation algorithm and high data dimensionality.

In terms of prediction overhead, Figure 4 shows that LDA has the most efficient prediction phase among all models and costs only 0.11 sec and 0.3J for prediction on a 2000 image

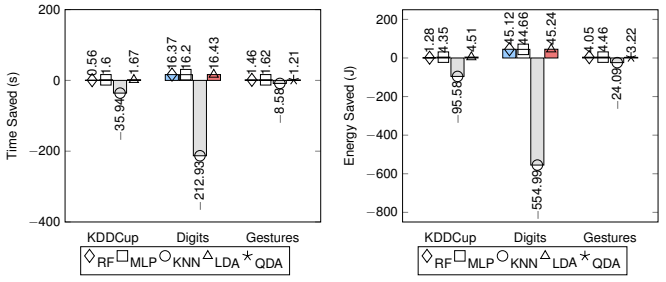

Fig. 5: RF, MLP, and LDA all demonstrate overhead savings when applied to data before transmission; most notably, RF applied to the Digits dataset saves 45.119 J and 16.37 seconds per 2000 image buffer. KNN is a poor choice for real-time anomaly detection scenarios given that it causes net time and energy loss across all datasets.

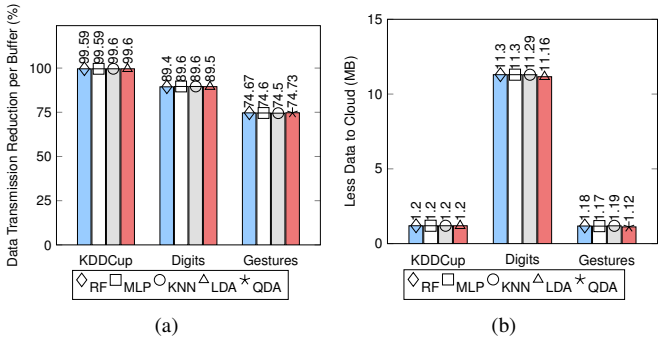

Fig. 6: From (a) we observe that the models reduce approximately 90% of the observations sent from Digits' buffers, 76% of the observations sent from Gestures' buffers, and 99.5% of the observations sent from KDDCup's buffers. The resulting saved cloud storage shown in (b) is substantial, especially for Digits dataset where we saved nearly 12 MB of data sent per 2000 image buffer.

buffer. It is followed by RF, MLP, and KNN in order. KNN is very expensive for prediction and costs nearly 230 seconds and 600J. This is because the distance computation between K neighbors and all M test samples is costly for 784 dimensional data points. If we prioritize prediction accuracy much higher than prediction time, KNN offers the best precision and recall values at the expense of costly prediction overhead. If the primary objective is to minimize transmission time delay, LDA offers a robust solution at the expense of low precision (i.e., 83.18%). RF and MLP lie in the middle ground and are the two most well-rounded solutions for this scenario. Because RF also has significantly less training overhead, we propose that RF is the best anomaly detection method for this image classification scenario.

3) Gestures Dataset: For this 64 feature dataset, all the models are trained using 3000 hand gesture observations out of a total of 12000 hand gesture observations with a 25.3% anomaly rate. This dataset has the second highest dimensionality and the highest anomaly rate among all datasets. In this scenario, the edge device aggregates 3000 observations (gestures) per buffer for the prediction phase and subsequent transmission. Note that we use QDA only on this dataset instead of LDA as LDA posted less than 20% precision. This is because LDA could only generate a linear fit for certain features in this dataset that exhibited quadratic patterns. Nevertheless, we benchmarked LDA and observed that it takes 0.43 seconds for the training phase and 0.051 seconds for the prediction phase. Therefore, if LDA demonstrated acceptable prediction accuracy for Gestures, it would have claimed the second most efficient training phase and the most efficient prediction phase among all models.

With regard to prediction accuracy, Figure 2 shows that QDA posts the highest recall out of all models at 98.2%. RF also performs well with 93.03% recall and 93.97% precision. It is followed by MLP and KNN in order. Given the overall high level of accuracy in anomaly detection, we assert that the edge device can save around 75% of the data (i.e., approximately 1.15 MB) per buffer sent to the cloud.

KNN training, as shown in Figure 3, is the most efficient and MLP training is the least efficient. RF training costs substantially more overhead than QDA. This is explained by the computation involved in creating a set of randomized decision trees for a pool of 64 features.

As far as prediction overhead, MLP has the most efficient prediction phase for the Gestures dataset and consumes 0.054 seconds and 0.142J per buffer. This is 274% more time efficient and 294% more energy efficient than the next best RF prediction phase. QDA and KNN follow in order. MLP is the clear choice for this scenario because it offers fast anomaly detection and the best cost-savings among all four models.

4) General Observations: We note that time and energy savings for each anomaly detection model is calculated by subtracting both prediction phase overhead and anomalous buffer transmission overhead from full buffer transmission overhead. Figure 5 shows the overhead savings observed when applying each model on each dataset. LDA provides the most data transmission overhead savings among all models. MLP, QDA, RF, and KNN follow in order. Considering that the deployment of KNN leads to net overhead loss across all benchmarked scenarios, its use may be eliminated from real-time anomaly detection scenarios.

Comparing both Figure 5 and Figure 4, we note that the rank ordering of models' overhead savings from least to greatest matches the rank ordering of models' prediction phase overhead from greatest to least. From Figure 4, note that QDA, which is used in place of LDA for the Gestures dataset, substantially exceeds LDA in prediction phase overhead consumption. This explains why MLP has the fastest prediction phase for the Gestures dataset but has the second and third slowest prediction phase for KDDCup and Digits datasets respectively. Also note that RF has the second most efficient prediction phase for Digits but is substantially less efficient than MLP for KDDCup. This suggests that RFs are more sensitive to testing buffer size than dimensionality. For reference purposes, the rank ordering of training time complexity from greatest to least is generally MLP, RF, QDA, LDA, and KNN across all datasets. Note that all rankings presented are based on testing with Python's scikit-learn library and experimental results may vary when using other software implementations

TABLE V: Sample use cases for different models.

Model	Suitable Applications
RF	Minimal training samples Delay sensitive applications
MLP	Non-linear relationship between training inputs and outputs Extremely time-sensitive application
KNN	Resource constrained training Delay insensitive application
DA	Resource constrained training and prediction Delay sensitive and mission critical application

of the machine learning models.

Figure 6 shows percentages of buffer size reduction and fewer MBs of data sent to the cloud upon applying each model on each dataset. The amount of data storage conserved on the cloud depends on the dateset's anomaly rate and dimensionality. Low anomaly rate indicates that there will be a proportionally smaller number of observations sent to the cloud. Low dimensionality indicates that there will be fewer bytes of data per observation sent to the cloud. The cloud would benefit the most when applying anomaly detection for a scenario with very high dimensional data and low anomaly detection rate. In this way, the cloud will not receive the vast majority of data points sent from the edge. With this in mind, we will examine the cloud storage savings observed for each discussed scenario. KDDCup has the lowest dimensionality (i.e., 3) and lowest anomaly rate (i.e., 0.35%) among all the datasets. Therefore, we are able to filter over 99% of the observations captured on the edge but save only 3 features per observation. This explains why we are only able to save approximately 1.2 MB sent to the cloud during prediction phase even though the buffer size is 50000. Digits has the highest dimensionality (i.e., 784) among all datasets and falls between the other two datasets with regard to anomaly rate (i.e., 10%). Despite having a considerably higher anomaly rate than KDDCup, Digit's dimensionality ensures that we save 784 data points sent to the cloud per observation. Thus, by filtering 90% of each Digits buffer we save roughly 12.3 MB sent to the cloud. Lastly, Gestures has the highest anomaly detection rate (i.e., 25.3%) and ranks second as far as dimensionality (i.e., 64). This indicates that there is a substantially higher proportion of observations needed to be notified to the cloud compared to the other scenarios and a moderate level of data points saved per filtered observation. These two factors slightly downgrade cloud storage savings and result in approximately 1.15 MB saved by the cloud per data buffer.

Based on the above results, Table V summarizes the scenarios best suited for each model.

V. Conclusion

In this paper, we explored the use of anomaly detection as an impactful edge mining technique in different IoT scenarios. We proved that RF, MLP, LDA, and QDA anomaly detection models have considerable potential to save edge device transmission overhead as well as cloud storage. The overhead-savings for a generic IoT scenario varies depending upon the anomaly rate and dimensionality of the transmitted data. We conclude that LDA has the most cost-effective anomaly detection phase that we have benchmarked across all scenarios and should be the primary choice for an extremely resource constrained edge device. QDA also has a very efficient anomaly detection phase but undoubtedly demands more overhead than LDA for the quadratic fit operation. We also conclude that RF is the most well-rounded anomaly detection method among all models featuring a comparably lightweight prediction phase and offering exceptional precision and recall. MLP also works very well for time-critical prediction tasks given that there are not significant resource constraints for training. The KNN classifier, despite its reliable prediction accuracy, demands excessive amounts of time and energy for anomaly detection, which rules out its use case in most IoT scenarios. The only reason to consider using KNN is in a case of stringent resource constraints for model training.

This work clearly demonstrates the overhead-savings potential of machine learning based anomaly detection on both edge and cloud. We also have provided a comprehensive overview of the tradeoffs involved in the deployment of these models. For future work, we aim to benchmark unsupervised classification methods for anomaly detection. Unsupervised machine learning methods are very useful when we do not have ground truth labeling but can infer properties from the training dataset. For example, Elliptical Envelope is a suitable technique for a dataset which expresses a multivariate gaussian distribution and an Isolation Forest is optimal for a dataset which expresses a multimodal distribution. For future contribution, we also aim to scale our experimental setup to other IoT platforms such as Cypress CYW43907.

References

[1] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive non-orthogonal multiple access for cellular iot: Potentials and limitations," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 55–61, 2017.

[2] Y.-L. Tsou, H.-M. Chu, C. Li, and S.-W. Yang, "Robust distributed anomaly detection using optimal weighted one-class random forests," in *IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1272–1277.

[3] E. I. Gaura, J. Brusey, M. Allen, R. Wilkins, D. Goldsmith, and R. Rednic, "Edge mining the internet of things," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3816–3825, 2013.

[4] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[5] K. Bhargava and S. Ivanov, "Collaborative edge mining for predicting heat stress in dairy cattle," in *2016 Wireless Days (WD)*. IEEE, 2016, pp. 1–6.

[6] I. Butun, B. Kantarci, and M. Erol-Kantarci, "Anomaly detection and privacy preservation in cloud-centric internet of things," in *IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 2610–2615.

[7] S. Bin, L. Yuan, and W. Xiaoyi, "Research on data mining models for the internet of things," in *International Conference on Image Analysis and Signal Processing*. IEEE, 2010, pp. 127–132.

[8] F. Giannoni, M. Mancini, and F. Marinelli, "Anomaly detection models for iot time series data," *arXiv preprint arXiv:1812.00890*, 2018.

[9] S. A. Aljawarneh and R. Vangipuram, "Garuda: Gaussian dissimilarity measure for feature representation and anomaly detection in internet of things," *The Journal of Supercomputing*, pp. 1–38, 2018.

[10] H. Haddad Pajouh, R. Javidan, R. Khayami, D. Ali, and K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in iot backbone networks," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2019.

[11] S. Raza, L. Wallgren, and T. Voigt, "Svelte: Real-time intrusion detection in the internet of things," *Ad hoc networks*, vol. 11, no. 8, pp. 2661–2674, 2013.

[12] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *IEEE Symposium on Security and Privacy*, May 2010, pp. 305–316.
[13] J. Zhang and M. Zulkernine, "Anomaly based network intrusion detection with unsupervised outlier detection," in *IEEE International Conference on Communications*, vol. 5, June 2006, pp. 2388–2393.
[14] G. Kotani and Y. Sekiya, "Unsupervised scanning behavior detection based on distribution of network traffic features using robust autoencoders," in *IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov 2018, pp. 35–38.
[15] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: a comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.
[16] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, "Iot healthcare analytics: The importance of anomaly detection," in *IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, March 2016, pp. 994–997.
[17] R. K. Mishra and R. Pandey, "Aspects of network architecture for remote healthcare systems," in *2nd International Conference on Computational Intelligence and Networks (CINE)*, Jan 2016, pp. 47–53.
[18] D. Stiawan, M. Y. Idris, R. F. Malik, S. Nurmaini, and R. Budiarto, "Anomaly detection and monitoring in internet of things communication," in *8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Oct 2016, pp. 1–4.
[19] T. Luo and S. G. Nagarajan, "Distributed anomaly detection using autoencoder neural networks in wsn for iot," in *IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.
[20] H. Sedjelmaci, S. M. Senouci, and T. Taleb, "An accurate security game for low-resource iot devices," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9381–9393, 2017.
[21] L. Lyu, J. Jin, S. Rajasegarar, X. He, and M. Palaniswami, "Fog-empowered anomaly detection in iot using hyperellipsoidal clustering," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1174–1184, 2017.
[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[23] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[25] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *International Journal of Engineering Research and Applications*, vol. 3, no. 5, pp. 605–610, 2013.
[26] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.
[27] B. Jiang, X. Wang, and C. Leng, "A direct approach for sparse quadratic discriminant analysis," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1098–1134, 2018.
[28] "Http (kddcup99) dataset," http://odds.cs.stonybrook.edu/http-kddcup99-dataset/, accessed: 2019-01-30.
[29] "Mnist digits," http://yann.lecun.com/exdb/mnist/, accessed: 2019-01-30.
[30] "Classify gestures by reading muscle activity," https://www.kaggle.com/kyr7plus/emg-4/metadata, accessed: 2019-01-30.
[31] B. Dezfouli, I. Amirtharaj, and C.-C. C. Li, "Empiot: An energy measurement platform for wireless iot devices," *Journal of Network and Computer Applications*, vol. 121, pp. 135–148, 2018.
[32] E. Upton and G. Halfacree, *Raspberry Pi user guide*. John Wiley & Sons, 2014.
[33] E. I. Gaura, J. Brusey, M. Allen, R. Wilkins, D. Goldsmith, and R. Rednic, "Edge mining the internet of things," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3816–3825, Oct 2013.
[34] S. M. Omohundro, *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.
[35] H. Neeb and C. Kurrus, "Distributed k-nearest neighbors," 2016.
[36] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, Dec 2016.
[37] D. H. Summerville, K. M. Zach, and Y. Chen, "Ultra-lightweight deep packet anomaly detection for internet of things devices," in *IEEE 34th International Performance Computing and Communications Conference (IPCCC)*, Dec 2015, pp. 1–8.
[38] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.
[39] G. Puskorius and L. Feldkamp, "Truncated backpropagation through time and kalman filter training for neurocontrol," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 4. IEEE, 1994, pp. 2488–2493.
[40] E. D. Karnin, "A simple procedure for pruning back-propagation trained neural networks," *IEEE transactions on neural networks*, vol. 1, no. 2, pp. 239–242, 1990.
[41] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.
[42] N. K. Thanigaivelan, E. Nigussie, R. K. Kanth, S. Virtanen, and J. Isoaho, "Distributed internal anomaly detection system for internet-of-things," in *13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2016, pp. 319–320.
[43] A. Saeed, A. Ahmadinia, A. Javed, and H. Larijani, "Intelligent intrusion detection in low-power iots," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 4, p. 27, 2016.