frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 06 January 2025
DOI 10.3389/fcomp.2024.1493399

OPEN ACCESS

EDITED BY
Yong Wang,
Guilin University of Electronic Technology, China

REVIEWED BY
Luca Deri,
University of Pisa, Italy
Sabina Rossi,
Ca' Foscari University of Venice, Italy

*CORRESPONDENCE
Sean Choi
✉ sean.choi@scu.edu

RECEIVED 09 September 2024
ACCEPTED 04 December 2024
PUBLISHED 06 January 2025

CITATION
Kapoor R, Anastasiu DC and Choi S (2025) ML-NIC: accelerating machine learning inference using smart network interface cards. *Front. Comput. Sci.* 6:1493399. doi: 10.3389/fcomp.2024.1493399

ML-NIC: accelerating machine learning inference using smart network interface cards

Raghav Kapoor[1], David C. Anastasiu[2] and Sean Choi[1]*

[1]Cloud Lab, Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, United States, [2]Anastasiu Lab, Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA, United States

Low-latency inference for machine learning models is increasingly becoming a necessary requirement, as these models are used in mission-critical applications such as autonomous driving, military defense (e.g., target recognition), and network traffic analysis. A widely studied and used technique to overcome this challenge is to offload some or all parts of the inference tasks onto specialized hardware such as graphic processing units. More recently, offloading machine learning inference onto programmable network devices, such as programmable network interface cards or a programmable switch, is gaining interest from both industry and academia, especially due to the latency reduction and computational benefits of performing inference directly on the data plane where the network packets are processed. Yet, current approaches are relatively limited in scope, and there is a need to develop more general approaches for mapping offloading machine learning models onto programmable network devices. To fulfill such a need, this work introduces a novel framework, called ML-NIC, for deploying trained machine learning models onto programmable network devices' data planes. ML-NIC deploys models directly into the computational cores of the devices to efficiently leverage the inherent parallelism capabilities of network devices, thus providing huge latency and throughput gains. Our experiments show that ML-NIC reduced inference latency by at least 6× on average and in the 99th percentile and increased throughput by at least 16*x* with little to no degradation in model effectiveness compared to the existing CPU solutions. In addition, ML-NIC can provide tighter guaranteed latency bounds in the presence of other network traffic with shorter tail latencies. Furthermore, ML-NIC reduces CPU and host server RAM utilization by 6.65% and 320.80 MB. Finally, ML-NIC can handle machine learning models that are 2.25× larger than the current state-of-the-art network device offloading approaches.

KEYWORDS

machine learning, SmartNIC, Netronome, data plane, inference

1 Introduction

Machine learning (ML) permeates a vast amount of everyday life, from personalized recommendations to stock market analysis and novel drug synthesis. While the machine learning models created to solve problems in these various fields are proven to be highly effective, these models often need large amount of time to make predictions (also referred to as model inference) on data instances. Often times, such limitation becomes a huge limiting factor for deploying ML models for latency critical applications. For example, applications such as high-frequency trading, military target recognition, pilots traveling at aircraft speeds (i.e., at least 621 mph) need to perform ML inference with the tight

latency budget of at most 50 ms. Making things worse, it is often not computationally and physically feasible to host large, effective ML models on directly on the devices like military aircraft. Thus, the ML model computations are often offloaded onto ground stations or edge devices, where data transmission can significantly add to the latency to execute model inference.

While many different types of machine learning accelerators have been developed, such as Graphical Processing Units (GPU) (Choquette et al., 2018), Field Programmable Gate Arrays (FPGA) (Fowers et al., 2018; He et al., 2018; Tong et al., 2017), and specialized application-specific integrated circuits (ASIC) (Chen et al., 2014; Jouppi et al., 2017), their efficiency is lowered by the data transfer time over the PCIe bus from the host system's network interface card (NIC). To overcome this challenge, we investigated methods to perform ML inference at the edge of the network to reduce the need and the overhead of transferring data from the edge to these accelerators. To achieve this, we note that the emergence of programmable data planes makes network devices (i.e., programmable switches and NICs) potential candidates for accelerating ML inference, especially given that programmable network devices have already been shown to be significantly power efficient while also providing high throughput and low latency in a variety of in-network computing tasks such as caching (Jin et al., 2017), consensus (Dang et al., 2020), and network monitoring (Kim et al., 2015). However, leveraging programmable data planes for machine learning inference still is an ongoing area of research with room for improvement.

Much of the prior works in this area has shown that network devices with programmable data planes, primarily programmable switches, demonstrate superior latency performance with minor degradation in model effectiveness (Zhang et al., 2023). While the line rate performance of programmable switches is beneficial for model inference, their limitation to match+action logic, memory size, cost, and the placement in the network restrict the feasibility and accuracy of models that can be mapped onto them. For example, since many modern machine learning algorithms rely on operations such as multiplication during inference, finite-sized match+action tables cannot support every possible combination of multiplied values. Even though prior methods have found ways around this, it was not without loss in model effectiveness. And as machine learning models continue to grow, additional losses in model effectiveness seem likely. In addition, prior methods focused on general models that may not be used in the real-world for low-latency applications.

To compensate for this limitation, we propose the Smart Network Interface Cards (SmartNICs) as a viable alternative. SmartNICs possess additional computational resources and several packet processing accelerators that can be adapted to mimic essential machine learning inference operations, such as multiplication and logarithm functions, more accurately. Furthermore, SmartNICs are much more cost and power efficient, are more easier to deploy and test.

Therefore, in this paper, we present ML-NIC, a framework for compiling and deploying trained machine learning models onto SmartNICs by providing intelligent model mapping methods. This current work mainly focuses on mapping tree-based models onto SmartNICs due to it's wide usage of low-latency applications, but we have proposals for future work with proposed methods to support inference for other types machine learning models as well. Compared to many prior works that implement machine learning algorithms onto programmable network devices, ML-NIC implementation uses more device parallelism in the inference process. Finally, our Python implementation of ML-NIC is made publicly available upon publication of the manuscript.

Our contributions include:

1. We present an algorithm to extract logic learned by a generic decision tree to facilitate parallelized feature analysis during inference.
2. We present a method to map and compile trained decision trees onto a SmartNIC in a manner that leverages its parallelism capabilities.
3. We demonstrate our framework's potential for accelerating the inference of decision trees for different tasks compared to conventional CPU and current state-of-the-art SmartNIC model deployment strategies.
4. We created an open-source project that contains all of ML-NIC's implementation and experimentation.[1]

The rest of this paper is organized as follows. Section 2 presents some background information on SmartNICs relevant to our work. Section 3 explains our approach toward deploying machine learning models onto a SmartNIC. Sections 4, 5, and 6 describe our experimental setup and discuss our results. Section 7 presents an overview of recent work that utilizes programmable data planes to accelerate the inference time for various machine learning algorithms on different problems. Section 8 points out future directions, and Section 9 ends with concluding remarks.

2 Background

In this section, we provide the background for ML-NIC and some underlying motivations.

2.1 SmartNIC

Smart Network Interface Cards (SmartNICs) possess additional computational resources and memory storage compared to traditional network interface cards. These resources enable SmartNICs to perform deep packet inspection, network function virtualization, and zero-trust security (Netronome Systems, 2024). As a result, offloading such operations to the SmartNIC frees a host system's CPU from conducting them. Compared to programmable switches, SmartNICs have more computational resources that can be leveraged. This motivates our choice to use SmartNICs for machine learning inference, since this process can be quite computationally intensive. For the rest of the section, we will focus on one particular type of SmartNIC: ASIC-Based Netronome SmartNICs supporting the NFP4000 architecture.

1 The project can be found on https://github.com/The-Cloud-Lab/ML-NIC.

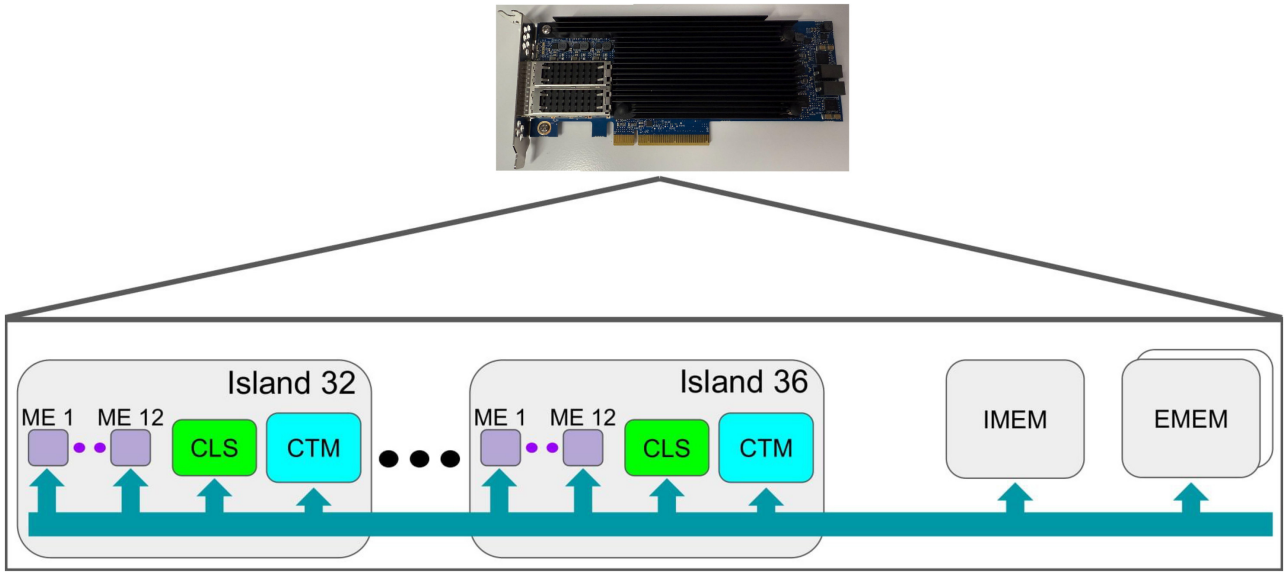

FIGURE 1
In this figure, we present a high-level overview of the NFP4000 architecture, focusing on the components most pertinent to ML-NIC.

2.1.1 ASIC-based Netronome SmartNIC

ASIC-based SmartNIC represents a type of SmartNIC where the ASIC is custom built to support programmability using a set of custom languages. One notable example of an ASIC-based SmartNIC is a Netromome SmartNIC that can be programmed using langauged called P4 and Micro-C. This project utilizes a specific subset of the Netromome SmartNICs, which support the NFP4000 architecture. To elaborate further, Netronome SmartNICs support the NFP4000 architecture feature 48 packet processing cores and 60 programmable flow processing cores (Corigine, 2020). In much of the literature and technical documentation, the flow-processing cores are referred to as microengines (ME), which we denote as purple squares in Figure 1. Each microengine acts as an independent 32-bit processor with its own code store and local memory to run different programs in parallel with the other microengines. The microengines can be programmed using a low-level language like Micro-C, an extended subset of C89, or a high-level language like P4. A key difference between Micro-C and P4 is that Micro-C provides the flexibility to program each microengine differently, whereas P4 defaults to loading the same program onto all microengines. However, both languages lack floating-point number support. We provide a high-level illustration of the NFP4000 architecture in Figure 1.

Each microengine supports 8 threads, where each thread runs the same program and has its own block of memory/registers. The following memory in a microengine is evenly partitioned among the 8 threads in a microengine:

- 256, 32-bit General-Purpose Registers—used for general per-packet computations.
- 256, 32-bit Transfer Registers—used for transferring data between memory regions.
- 128, 32-bit Next-Neighbor Registers—used for communicating between neighboring microengines in the same island.
- 4kB of Local Memory—used for additional data storage as needed.
- 120 Signal Registers—used to notify threads that a certain hardware event has occurred.

The partitioning of memory among the 8 threads facilitates fast context switching between them, so they can process different packets efficiently (Siracusano et al., 2022).

The microengines are organized into islands. While these islands can vary in number of microengines and specialized functionality, standard islands, shown in Figure 1, contain 12 microengines with two regions of memory shared between all the microengines in the island: Cluster Local Scratch (CLS) and Cluster Target Memory (CTM). CLS, denoted in green in Figure 1, commonly stores small forwarding tables shared between the microengines (Wray, 2014). CTM, denoted in cyan in Figure 1, holds packet headers and coordinates between the microengines and other subsystems on the card (Wray, 2014). As CTM is larger than CLS, more clock cycles are required to read/write to CTM.

Outside of the islands, the Netronome SmartNICs have three additional memory units, as shown in Figure 1, shared with all microengines: one Internal Memory Unit (IMEM) and two External Memory Units (EMEM) (Langlet, 2019). IMEM is used for storing packet payloads and medium-sized match-action tables (Wray, 2014). EMEM is used to store larger match-action tables and other flow statistics (Wray, 2014). As these three memory units are the largest of those mentioned prior, with EMEM being larger than IMEM, they require a greater number of clock cycles to read/write to them. Microengines can access data in all these

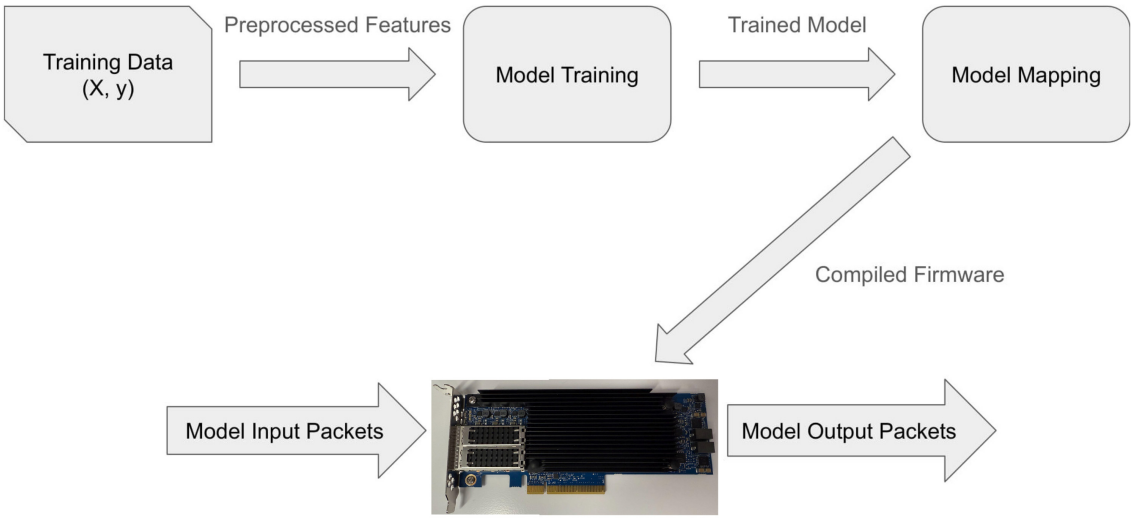

FIGURE 2
This figure shows the big picture of the ML-NIC architecture.

memory regions using the Command Push Pull (CPP) bus, denoted in turquoise in Figure 1.

Based on our knowledge of data access times for the different memory regions, hardware signals, and transfer registers, we design ML-NIC to efficiently leverage these resources to ensure a high degree of performance from a SmartNIC.

3 ML-NIC architecture

ML-NIC comprises three components: machine learning model training, model mapping, and model deployment, as shown in Figure 2. We explain the details of each component below.

3.1 Model training

In the machine learning model training component, we consider a labeled dataset (X, y), where $X \in \mathbb{R}^{mxn}$ represents our data matrix (m data points, n features) and $y \in \{1 \ldots q\}^m$ represents our class labels (q possible classes). We make no assumptions on whether X consists of only continuous features, only categorical features, or a mix. We assume the continuous features are normalized within the range [0, 1]. We do so to simplify the range of numerical representation that the first iteration of ML-NIC needs to account for. We find that this assumption is reasonable, since data normalization is a common technique in machine learning to prevent certain features from dominating over other features due to differences in scaling. For the categorical features, we assume that they are one-hot encoded (Liu, 2017) (i.e., a feature with three categories is expanded to three features with values 0 or 1). This dataset is used to learn the parameters of a particular machine learning model.

In this first iteration of ML-NIC, we choose to focus on decision trees for three reasons. First, we cite the relative computational simplicity of tree traversal compared to floating-point operations in hardware without a Floating-Point Unit (FPU) like a SmartNIC that was designed for fast computations at network line rates.

Second, in comparison to other classical supervised machine learning models, such as Naive Bayes, k-nearest neighbor, and support vector machine, we find that the decision tree is the more suitable choice for offloading. With Naive Bayes, it is known that the algorithm perform poorly when the features used for training are not conditionally independent. However, in practice, decision trees can perform well even if the features are correlated. For k-nearest neighbor, we find that the required storage of every training instance to be an obstacle for offloading onto network devices, especially given the size of modern datasets. Even if only a selection of the training set was used to make offloading feasible, this could result in more significant performance degradation in certain machine learning problems. With respect to support vector machines for multiclass classification, the model may not be optimal for offloading with a large number of features and number of classes. For this explanation, we temporarily denote m to be the number of features, q as the number of classes (greater than 2 for multiclass classification), and z as the number of support vectors. First, assume the support vectors for the support vector machine can be stored and there exists suitable means of multiplication and a kernel on an off-the-shelf programmable network device that are at least as expensive as a comparison operation. Since off-the-shelf programmable network device are optimized for match + action, we expect the device architecture to have an efficient compare operation. Then, we see that the support vector machine requires at least $m \times z \times q$ multiplications, whereas a decision tree would require at most m compare operations for a single inference. Therefore, based on instruction count, the decision tree model has a better chance of yielding inference latency reduction when offloaded onto an off-the-shelf programmable network device.

Third, we note that tree-like machine learning models, such as XGBoost (Chen and Guestrin, 2016), are commonly used to learn

```
Require: S is a valid subset of the training dataset D, R is a set of stopping criteria for the algorithm, M is a valid impurity metric to locally optimize, split_node is a valid function for a splitting S at a node

function train_tree(S, R, split_node)
    c = majority_label(S)
    tree_node = Node(label = c)
    if not_satisfied(R) then
        ms = list()
        splits = split_node(S)
        for each split ∈ splits do
            ms.append(M(split))
        end for
        best_split = splits[arg_optimal(ms, M)]
        for each s ∈ best_split do
            tree_node.insert_branch(train_tree(s, R, split_node))
        end for
    end if
    return tree_node
end function
```

Algorithm 1. Decision tree training algorithm.

tasks from structured tabular data over neural networks given faster training time, potential performance gain, and model transparency. We use the decision tree model to show that our framework can be used in real-world applications and offloading a machine learning model onto a SmartNIC can reduce model inference latency significantly. While our focus is currently on decision tree inference, we provide a discussion on how our framework can be augmented to account for additional machine learning models in Section 6.

To train a decision tree, we consider the high-level algorithm outlined in Algorithm 1. Since finding the globally-optimal tree structure for a learning task is computationally challenging, locally-optimal heuristic algorithms are used such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1986), and CART (Breiman et al., 1984). In practice, metric M is commonly Information Gain in the case of ID3 and C4.5 or Gini Impurity for CART. Formulations for these metrics are provided in Equations 1, 2. In Equations 1, 2, we further define C as the number of classes, P as the number of splits, S_c as the number of examples in the training dataset subset with class label c, and $S_{p,c}$ as the number of examples in the pth split of the training dataset subset with class label c.

$$Gini(S) = \sum_{c=1}^{C} \frac{|S_c|}{|S|}\left(1 - \frac{|S_c|}{|S|}\right) \tag{1}$$

$$Info(S) = \sum_{c=1}^{C}\left(-\frac{|S_c|}{|S|}\log_2\left(\frac{|S_c|}{|S|}\right)\right) - \sum_{p=1}^{P}\frac{|S_p|}{|S|} \sum_{c=1}^{C}\left(-\frac{|S_{p,c}|}{|S_p|}\log_2\left(\frac{|S_{p,c}|}{|S_p|}\right)\right) \tag{2}$$

3.2 Model mapping

Before discussing the technical details of decision tree mapping onto a SmartNIC, we discuss our mapping approach at a high-level. To run inference, we find the disjunctive normal form (Roth, 2016) of a decision tree. In the disjunctive normal form, the logic for assigning a class label to a data instance is expressed as a disjunction of conjunctions [i.e., (condition 1 and condition 2 and ...) or (condition 3 and condition 1 and ...) or ...]. Each conjunction in the disjunction (i.e., condition 1 and condition 2 and ...) represents a path from the root node to a leaf node in a decision tree. We prefer the disjunctive norm form over the typical tree structure of a decision tree for inference, since it makes executing inference in a parallelized manner more convenient. To parallelize the inference process from the disjunctive normal form, we take the conditions from all the conjunctions that correspond to a particular feature, noting which path in the decision tree the condition corresponds to. To run inference on a data instance then, the conditions for each feature can be evaluated in parallel, where the result of each feature evaluation yields a set of paths in the decision tree that are possible for the data instance to take. Then, by aggregating the all possible paths and taking the intersection among them, a single path can be found. By matching the path to its corresponding class label, the decision tree inference process is complete.

To map a decision tree onto a SmartNIC, we take the output of the machine learning model training process (i.e., a pickle file) and proceed to generate an implementation of the SmartNIC data plane. Currently, we support SmartNICs that are programmable in Micro-C, primarily SoC-based Netronome SmartNICs. In the current iteration of our framework, we consider a trained decision tree classifier C with l leaf nodes, where l is at most 256. We make no additional assumptions on the number of splits per non-leaf node or the training algorithm used. Based on the number of leaf nodes l in model C and number of features n in X, there are three possible scenarios for mapping C onto the SmartNIC:

- The model can fit on one island of the SmartNIC. Each island is then programmed with its own set of feature computation, result aggregation, and packet collection microengines (i.e., inference for model C is run on all the islands).
- The model can fit on the entire SmartNIC with one feature assigned per feature microengine and one packet collection microengine.
- The model can fit on the entire SmartNIC with multiple features assigned per feature microengine and one packet collection microengine.

After selecting one of the three above mapping schemes, the next step is to extract the logic (i.e., find disjunctive normal form and extract the conditions that match to a particular feature) learned by model C. To do so, we iterate through all the n features in X and perform a depth-first search through the decision tree. We record the operation for those nodes that run a comparison operation on our feature of interest and continue the depth-first search until all the leaf nodes have been reached. Formally, Algorithm 2 illustrates our logic extraction approach.

```
Require: node points to valid node in decision tree, ftre is feature seen by decision tree during training, clt has enough space to store decision tree logic for ftre

function get_logic(node, ftre, clt)
    if is_leaf(node) then
        clt.insert(node.prediction)
    else
        if node.ftre = ftre then
            clt.insert(node.logic)
        end if
        for each child ∈ node.children do
            clt.insert(get_logic(child, ftre, clt))
        end for
    end if
    return clt
end function
```

Algorithm 2. Decision tree logic extraction algorithm.

We also assign the each of microengines on the SmartNIC as one of three types: packet collection, feature computation, and result aggregation. The packet collection microengine(s) are programmed to signal the CTM packet engine that they are ready to receive packets. Once a packet is received, the packet collection microengine(s) will verify that packet is a model input packet, extract the features from the packet payload, and asynchronously signal all the feature computation microengines of the inference request and transmit the corresponding feature to each via transfer registers.

The feature computation microengines are responsible for evaluating the conditions on a feature for a given data instance and determine which paths in the decision tree are possible. Since each feature computation microengine is responsible for different features and run simulatenously, all the features can be evaluated and all the possible paths in the decision tree can be determined in parallel. To implement the conditions and determine the possible paths per feature on the SmartNIC, we use Micro-C if-statements to evaluate the conditions and update an array of integers to reflect which paths are possible. For the update, we treat the array of integers as a single bit string, where most significant bit in the integer at the last index in the array corresponds to path 1. We assign paths based on the order in which the nodes are encountered by Algorithm 2. Since the values for comparison in the conditions for evaluating each feature in the decision tree and the features themselves can be floating-point, and the SmartNIC does have an FPU, we consider an alternative floating-point representation. We represent floating-point numbers on the SmartNIC using a fixed-point representation that consists of 16 bits, where the last 13 bits represent the non-integer portion of a floating-point number. As each feature computation microengine completes its evaluation of its correponding feature, they notify the result aggregation microengine(s) of the decision tree paths that are possible based on the feature they each evaluation.

Once all the feature computation microengines finish their evaluation, the result aggregation microengine(s) finds the intersecting path the decision tree between all the possible paths, matches the path to the corresponding class label, and sends an asynchronous signal to the packet collection microengine(s) along with the class label via transfer register(s). Once the packet collection microengine(s) receives the signal from the result aggregation microengine(s), it edits the original packet payload with the class label for the data instance and notified the CTM packet engine that the packet needs to be transmitted. Also note that during the time the feature computation and result aggregation microengines are completing their tasks, the packet collection microengine(s) are editing the model input packet's header in preparation for transmission as a model output packet.

To program all the packet collection, feature computation, and result aggregation microengines, separate Micro-C code is written to program each microengine to complete their specific task for model inference, whereas prior methods often program all the microengines with one piece of P4 code to perform the same tasks for the model inference and do not fully leverage the parallel operating capacity of the SmartNIC. Example Micro-C code for packet collection, feature computation, and result aggregation can be found in Appendix Listings 1–3.

3.3 Model deployment

Once all the Micro-C code files are created, they are all compiled and linked to generate the device firmware to run on the data plane in the model deployment component. Then, the firmware file output is loaded onto the SmartNIC. An example of the full process is shown in Appendix Listing 4. Each microengine assumes a specified behavior based on one of the three microengine assignments specified above. The SmartNIC can now ingress packets with features in the packet payload, run machine learning inference in a parallelized manner, and egress packets with the classification result as the packet payload.

4 Experimental setup

4.1 Testbed

Our testbed consists of two Dell PowerEdge Rack Servers. Server 1 hosts an NVIDIA Mellanox Bluefield-2 DPU 25 GbE SmartNIC for packet transmission and data collection. Server 2 hosts a Netronome AgilioCX 2 × 25 GbE SmartNIC, on which our decision tree models are deployed. Both systems are directly connected via qsfp cable between the Mellanox and Netronome SmartNICs. We illustrate our setup in Figure 3.

4.2 Datasets and models

Our evaluation considers four tasks: land mine detection, satellite image pixel classification, gas sensor drift compensation, and network traffic classification. The main characteristics of the datasets used for each task can be found in Table 1. We

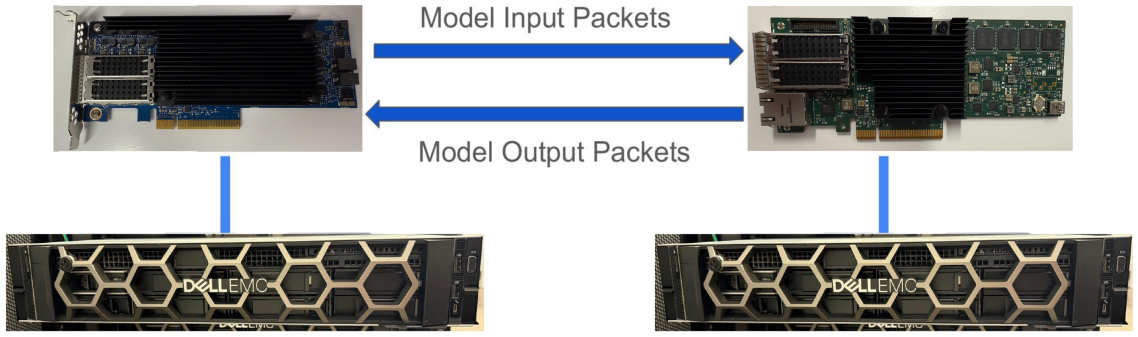

FIGURE 3
The testbed setup used for evaluation. The left server hosts the Netronome AgilioCX 2 × 25 GbE. The right server hosts the NVIDIA Mellanox Bluefield-2 DPU 25 GbE.

TABLE 1 Summary of datasets used (refer to dataset subsections for class label abbreviations).

Attribute	Mine	Landsat	Gas	CICIDS
# of features	3	36	128	7
# of data instances	338	6,435	13,910	22,887,218
# of classes	5	6	6	7
# of training data	270	4,435	11,128	500,000
# of test data	68	2,000	2,782	6,957,375

TABLE 2 Summary of decision tree models created.

Parameter	Mine	Landsat	Gas	CICIDS
# of leaves	114	256	256	89
Depth	17	15	24	15
# of nodes	227	511	511	177
Min samples leaf	1	1	1	2
Min samples split	2	2	2	2
Min impurity decrease	0	0	0	0.00001
Max leaf nodes	None	256	256	None

train a decision tree model for each task using the scikit-learn library (Pedregosa et al., 2011). We summarize the hyperparameters used for each tree in Table 2.

For hyperparameters not explicitly mentioned in the table that can be tuned for the decision tree models (i.e., criterion, splitter, max features, etc.), we resort to the default values provided by scikit-learn.

4.2.1 Dataset preprocessing

As mentioned in Section 3, we assume the continuous features are in the range [0, 1] and categorical features are one-hot-encoded. To achieve this, we apply min-max normalization to scale the continuous features of each dataset to range between 0 and 1 using the training set. Test features that lie outside the range [0, 1] after min-max normalization has been applied are clipped to the closest endpoint. We also one-hot-encode the categorical features for each dataset based on the values observed from the training set. If the categorical features in the test set take on values not observed in the training set, the one-hot-encoded feature is represented as a bit string of zeros.

4.2.2 Land mine detection

We use the Land Mines dataset (Yilmaz et al., 2018) for the land mine detection task. The authors propose three features to classify a mine into five types, Null, Anti-Tank, Anti-Personnel, Booby-Trapped Anti-Personnel, and M14 Anti-Personnel, with 65 − 71 samples per class. Our motivation for choosing this dataset is based on the number of features (8 after data preprocessing), where we can evaluate the first SmartNIC mapping scenario (fitting on one island) as described in Section 3. In later sections, we will refer to this dataset as Mine.

4.2.3 Satellite image pixel classification

We use the Statlog (Landsat Satellite) dataset (Srinivasan, 1993) for the satellite image pixel classification task. The goal of this task is to examine multispectral values from a 3 × 3 neighborhood of a satellite image and classify the central pixel as one of five classes: Red Soil, Cotton Crop, Gray Soil, Damp Gray Soil, Soil with Vegetation Stubble, Mixture, or Very Damp Gray Soil. There are 626 − 1,533 samples per class. Our motivation for choosing this dataset is based on the number of features (36), where we can evaluate the second mapping scenario (fitting on whole SmartNIC, one feature per microengine) as described in Section 3. In later sections, we will refer to this dataset as Landsat.

4.2.4 Gas sensor drift compensation

We use the Gas Sensor Array Drift dataset (Rodríguez-Luján et al., 2014) for the gas sensor drift compensation tasks. This dataset consists of measurements from 16 chemical sensors to identify six gases, Ammonia, Acetaldehyde, Acetone, Ethylene, Ethanol, and

Toluene, with 1,508 − 3,009 samples per class. Our motivation for choosing this dataset is based on the number of features (128), where we can evaluate the third mapping scenario (fitting on whole SmartNIC, multiple features per microengine) as described in Section 3. In later sections, we will refer to this dataset as Gas.

4.2.5 Network traffic classification

We use the CICIDS2017 dataset (Sharafaldin et al., 2018) for the network traffic classification use case. This use case's purpose is to identify network flows as benign or malicious (brute force attack, heartbleed attack, botnet, DoS attack, DDoS attack, web attack, infiltration attack). However, we follow the approach used by Xavier et al. (2021) to generate the dataset for classifying individual network packets rather than network flows and the training and tests sets based on the network flows instead of the conventional stratified 80/20 split used for the above datasets above. In the dataset, the packets were labeled as benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS Slowloris, Web Brute Force, or Port Scan. Each class has 30,059 − 20,121,944 samples. Our motivation for choosing this dataset is based on its use in the work by Xavier et al. (2021), which is similar to our approach. Our evaluation on this dataset clearly compares our approach and Xavier et al. (2021)'s approach. In later sections, we will refer to this dataset as CICIDS.

4.3 Baselines

We compare our approach against the following two baselines. First, we implement a traditional CPU baseline, which uses socket programming to receive incoming packets, extract the payload, run inference with the trained scikit-learn decision tree, and build and send a model output packet with the model prediction in the packet payload.

Second, we implement the approach developed by Xavier et al. (2021) using P4-16. Like our approach, Xavier et al. (2021)'s approach traverses through the scikit-learn decision tree structure, extracts the model's logic, and rebuilds the tree in P4 using Python. Note that the original implementation was solely created for the CICIDS dataset, and the authors did not provide a method to handle floating-point features. In evaluating this method on the other datasets, we modified it slightly to use our fixed-point representation of floating-point numbers. Also, due to limitations with Xavier et al. (2021)'s approach, we could not evaluate it on larger decision trees, such as those generated with the Landsat and Gas Datasets.

4.4 Evaluation metrics

In our experiments, we measure the effectiveness, (average and tail) latency, throughput, and hardware utilization of ML-NIC against the baselines.

For effectiveness, we measured the accuracy, F1 score, recall, and precision metrics on each dataset's test set. Given that our datasets are for multiclass classification tasks, we take the macro-average (i.e., unweighted mean) of the per-class scores for the F1 score, recall, and precision measurements.

For latency, we collected the time between model input packet transmission and model output packet reception on server 1 in microseconds for 1,000 packets. In addition to our vanilla latency experiments (i.e., no CPU load or network link utilization), we conduct latency experiments with background traffic on the network link and CPU load. We generate random network traffic at different speeds using Tcpreplay for latency experiments with background traffic to achieve 25%, 50%, and 99% network link utilization. We use stress-ng for latency experiments with CPU load and generate CPU loads of 25%, 50%, and 99%. To ensure an apples-to-apples comparison with the CPU baseline, we append zero padding to the model input packets for our approach and the P4 baseline. Hence, they are the same size as the CPU model input packets. Note that, when collecting the data for the CPU baseline, we remove the time taken to decode the data features and encode the model's prediction.

For throughput, we use Tcpreplay to loop through the PCAP files containing the test set packets for each dataset at top speed. Simultaneously, we also run Tshark to filter and collect the model prediction packets for 60 seconds.

Lastly, for hardware utilization, we run Tcpreplay for 60 seconds like we did for the throughput experiment and measure CPU, server host RAM, and SmartNIC memory utilization. We also measure CPU, server host RAM, and SmartNIC memory utilization 30 seconds before and after running Tcpreplay for reference. We do not explicitly measure SmartNIC microengine utilization in this experiment. Instead, we use the results from our network link utilization and CPU load latency experiments as a proxy for SmartNIC microengine utilization.

5 Results

From our experimental results, we demonstrate the following:

1. Our approach achieves effectiveness scores similar to those of the CPU baseline and identical to the P4 baseline.
2. Our approach has better a latency guarantee than the CPU and P4 baselines in various network link utilization and CPU load scenarios.
3. The throughput of our approach is significantly greater compared to the CPU baseline and on par with the P4 baseline.
4. Our approach uses fewer server host resources (i.e., CPU and server RAM) compared to the CPU and P4 baselines.

5.1 Effectiveness scores

As shown in Table 3, the effectiveness scores between our approach and the CPU baselines are similar with minor degradation. For conciseness, we only show the plots of accuracy and F1 score of the models, since the precision and recall results follow a similar pattern.

TABLE 3 Effectiveness measurements on all datasets.

Dataset	Measure	CPU	Xavier et al. (2021)	ML-NIC
Mine	Accuracy (%)	**58.82**	57.35	57.35
	F1 score (%)	**58.22**	56.92	56.92
Landsat	Accuracy	**85.65**	N/A	85.55
	F1 score (%)	**83.93**	N/A	83.76
Gas	Accuracy (%)	**97.41**	N/A	95.15
	F1 score (%)	**97.26**	N/A	94.78
CICIDS	Accuracy (%)	**95.12**	**95.12**	**95.12**
	F1 score (%)	**47.04**	**47.04**	**47.04**

Bold values indicates best values found for given metric during experiments.

For the Mine dataset, we note differences of 1.47%, 1.30%, 1.35%, and 1.54% across the accuracy, F1 score, precision, and recall metrics. For the Landsat dataset, we note differences of 0.100%, 0.17%, 0.14%, and 0.20% across the accuracy, F1 score, precision, and recall metrics. For the Gas dataset, we note differences of 2.26%, 2.49%, 2.00%, and 2.66% across the accuracy, F1 score, precision, and recall metrics. For the CICIDS dataset, our approach and the CPU baseline do not differ in accuracy, F1 score, precision, or recall. This is because all the features used to train the scikit-learn decision tree are integers, so no quantized representation of features is needed as with the previous three datasets. Our approach achieves identical effectiveness scores as the P4 baseline on the Mine and CICIDS datasets.

5.2 Latency

From the latency data we collected, we provide zoomed-in empirical cumulative distribution functions (eCDF) for each dataset, network link utilization, and CPU load in Figure 4. We also provide more concrete numbers on the 50th, 99th, and 99.9th percentiles across each dataset, link utilization, and CPU load in Tables 4, 5. From our experiments, we make the following observations. We generally see a significant gap in the latency measurements between ML-NIC and the CPU baseline and a very small gap between ML-NIC and Xavier et al. (2021)'s approach. Specifically, we found that ML-NIC's latency can be at least 132.62 μs faster than the CPU baseline and 1.35 μs faster than Xavier et al. (2021)'s approach in the 50th percentile. However, there is a significant difference in the tails between ML-NIC and Xavier et al. (2021)'s approach, suggesting that ML-NIC has a stronger latency guarantee. Based on the 99.9th percentiles, we see that the tail latency of Xavier et al. (2021)'s approach can be at least 1.53× larger than ML-NIC's tail.

Looking at impact of high network link utilization and CPU load, we observe very minimal fluctuation in the eCDFs of the ML-NIC and Xavier et al. (2021)'s approach. This suggests that both approaches are robust against high network link utilization and CPU load on these datasets. But, there is a more noticeable impact of high network link utilization and CPU load on the CPU baseline. As the network link utilization increases, we tend to see more probability mass shift toward the higher latency in the eCDF. With respect to the 99.9th latency percentile, we see an increase of at least 1.82× from 0% link utilization to 99% link utilization. Concerning the increases in CPU load, there is a more significant shift in the eCDF curves toward higher latencies. Referring to the 99.9th percentile, there is a latency increase of at least 20.27× from 0% CPU load to 99% CPU load.

5.3 Throughput

As seen in Figure 5, there is a significant improvement in throughput with our approach compared to the CPU baseline. In our approach, we note 24.80×, 19.30×, 16.95×, and 20.11× more packets per minute compared to the CPU baseline across the Mine, Landsat, Gas, and CICIDS datasets. Furthermore, our approach yields moderately higher throughput than the P4 baseline in the Mine and CICIDS dataset. In our approach, we observe 1.26× and 1.11× more packets per minute compared to the P4 baseline for the Mine and CICIDS datasets.

5.4 Hardware utilization

From our hardware utilization experiment, we report the minimum, maximum, and average CPU and server host RAM utilization in Table 6. We also report the SmartNIC memory utilization as a constant, since dynamic memory allocation is not available on the AgilioCX 2 × 25 GbE SmartNIC. Since we are not able to directly measure the SmartNIC RAM used for the CPU baselines, we approximate it. Our approximation takes an unweighted average of the ratio of size of SmartNIC firmware for the CPU baselines over the size of the SmartNIC firmware for the P4 baselines and our approach multiplied by the SmartNIC RAM used by those models for each of the datasets.

From Table 6, we see that ML-NIC consistently uses lower resources for average and maximum host system's CPU, host system's RAM, and SmartNIC's RAM usage compared to the CPU baseline and Xavier et al. (2021)'s method. In the case where there the CPU baseline has slightly lower minimum CPU usage than the ML-NIC on the CICIDS dataset, this measurement likely corresponds to some degree of randomness in the measurement, since ML-NIC also achieved the same minimum CPU usage on the Gas dataset. In addition to the CPU baseline having a higher maximum CPU usage than ML-NIC by at least 7.91×, we also observe that Xavier et al. (2021)'s method can achieve similar or higher levels of maximum CPU usage. We attribute this to the runtime environment (RTE) server that is running on our host system, which is needed to run P4 code. We believe this also accounts for slightly higher host system RAM usage. For the SmartNIC's RAM usage, we observe our proxy for the CPU baseline to be lower than Xavier et al. (2021)'s method. Since the firmware running on the SmartNIC for the CPU baseline runs as a regular NIC, the SmartNIC would not require additional memory beyond storing extracting packet headers into local memory or general purpose registers. Furthermore, the larger SmartNIC RAM usage from Xavier et al. (2021)'s method likely occurs because their

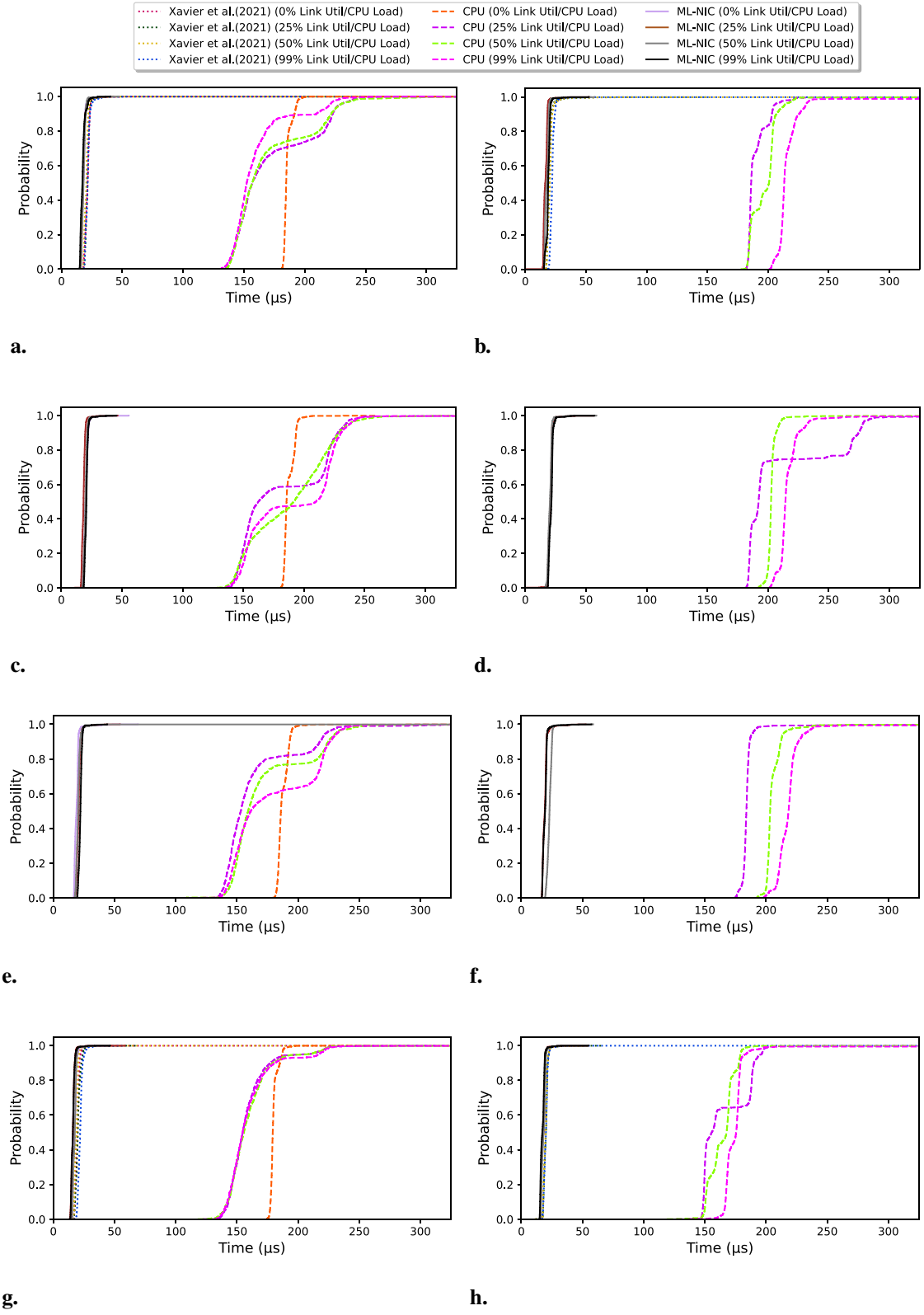

FIGURE 4
(A, C, E, G) Depict the eCDFs for the latency experiments conducted using Tcpreplay to saturate the network link on all the datasets. **(B, D, F, H)** Depict the eCDFs for the latency experiments conducted using stress-ng to generate a CPU load on all the datasets. **(A)** Mine Tcpreplay. **(B)** Mine stress-ng. **(C)** Landsat Tcpreplay. **(D)** Landsat stress-ng. **(E)** Gas Tcpreplay. **(F)** Gas stress-ng. **(G)** CICIDS Tcpreplay. **(H)** CICIDS stress-ng.

TABLE 4 Latency measurements using tcpreplay (μs).

Dataset	Link util (%)	Percentile	CPU	Xavier et al. (2021)	ML-NIC
Mine	0	50	184.77	21.55	**18.00**
		99	197.39	25.41	**23.60**
		99.9	202.94	71.26	**35.54**
	25	50	156.93	21.30	**17.73**
		99	243.95	27.35	**23.63**
		99.9	285.89	65.31	**36.11**
	50	50	156.96	20.41	**18.15**
		99	267.83	27.33	**21.53**
		99.9	364.59	60.37	**31.05**
	99	50	152.32	21.55	**17.24**
		99	230.96	31.10	**24.45**
		99.9	454.20	70.65	**36.48**
Landsat	0	50	185.61	N/A	**19.99**
		99	200.27	N/A	**24.88**
		99.9	260.64	N/A	**39.63**
	25	50	162.39	N/A	**18.27**
		99	247.78	N/A	**22.70**
		99.9	365.45	N/A	**36.65**
	50	50	191.58	N/A	**19.82**
		99	258.51	N/A	**23.87**
		99.9	458.64	N/A	**42.08**
	99	50	209.91	N/A	**20.63**
		99	247.88	N/A	**24.96**
		99.9	503.92	N/A	**37.40**
Gas	0	50	185.99	N/A	**18.91**
		99	198.86	N/A	**26.57**
		99.9	287.77	N/A	**48.31**
	25	50	153.22	N/A	**20.60**
		99	255.51	N/A	**27.00**
		99.9	367.69	N/A	**41.65**
	50	50	159.18	N/A	**20.13**
		99	251.86	N/A	**28.91**
		99.9	410.64	N/A	**45.20**
	99	50	161.33	N/A	**22.11**
		99	244.74	N/A	**25.35**
		99.9	523.63	N/A	**43.98**
CICIDS	0	50	179.30	19.58	**17.21**
		99	189.71	23.86	**21.66**
		99.9	198.83	54.91	**31.70**
	25	50	154.72	20.00	**15.97**
		99	224.22	26.66	**19.76**
		99.9	277.53	55.06	**29.29**
	50	50	155.85	18.94	**17.55**
		99	225.52	23.11	**20.68**
		99.9	442.80	55.48	**33.41**
	99	50	154.81	21.91	**15.99**
		99	226.99	28.21	**20.39**
		99.9	483.02	51.85	**33.80**

(Continued)

TABLE 4 (Continued)

Bold values indicates best values found for given metric during experiments.

approach involves running packet collect, feature computation, and result aggregation on every microengine. Since ML-NIC distributes these operations across multiple microengines, the resulting SmartNIC RAM usage would be lower by at least 46.05×.

6 Discussion

6.1 Generalization

The work focuses on converting trained scikit-learn decision trees into Micro-C for deployment onto a SmartNIC. We focused on the Netronome AgilioCX 2 × 25 GbE. Deployment across different SmartNICs (assuming Micro-C support) may require significant code changes to accommodate the resources available on the card compared to the baselines.

6.2 Benefits

Despite their resource constraints compared to a host system's CPU, SmartNICs show potential as alternative hardware for deploying appropriately sized decision tree models. Deploying the decision tree model onto the SmartNIC, which brings it closer to the network edge, saves latency time by removing the need to transfer data over the PCIe bus from the NIC to the CPU without additional hardware. Furthermore, the lower-level programming used in our approach compared to the P4 baseline allows us to leverage device parallelism to deploy larger decision trees.

6.3 Scope

Our work only considers deploying decision tree models trained for various tasks onto a SmartNIC. Improvements in any specific use case are beyond the scope of our work.

6.4 Limitations

Our method's limitations depend on the SmartNIC's memory, computational, asynchronous I/O, and data rate constraints.

TABLE 5 Latency measurements using stress-ng (μs).

Dataset	CPU load (%)	Percentile	CPU	Xavier et al. (2021)	ML-NIC
Mine	25	50	186.16	20.16	**16.45**
		99	221.97	31.82	**20.43**
		99.9	343.53	74.56	**32.35**
	50	50	200.85	20.12	**17.69**
		99	222.64	31.16	**20.96**
		99.9	535.06	67.87	**34.50**
	99	50	213.40	22.12	**18.84**
		99	332.46	27.96	**22.66**
		99.9	7,979.24	76.50	**36.86**
Landsat	25	50	192.30	N/A	**20.69**
		99	295.65	N/A	**24.48**
		99.9	405.26	N/A	**41.19**
	50	50	202.55	N/A	**20.23**
		99	215.20	N/A	**24.12**
		99.9	7,582.86	N/A	**38.09**
	99	50	214.34	N/A	**21.54**
		99	260.63	N/A	**25.80**
		99.9	9,008.55	N/A	**40.27**
Gas	25	50	183.90	N/A	**19.17**
		99	201.89	N/A	**27.42**
		99.9	469.44	N/A	**45.69**
	50	50	203.31	N/A	**23.01**
		99	238.13	N/A	**26.94**
		99.9	8,852.89	N/A	**46.95**
	99	50	218.32	N/A	**19.40**
		99	251.67	N/A	**25.54**
		99.9	16,182.94	N/A	**46.74**
CICIDS	25	50	155.58	20.13	**18.17**
		99	201.23	23.99	**21.31**
		99.9	357.74	51.48	**31.69**
	50	50	167.66	19.88	**18.18**
		99	184.74	24.39	**22.68**
		99.9	481.76	50.01	**30.33**
	99	50	175.56	20.15	**17.16**
		99	198.38	24.00	**21.83**
		99.9	4,030.55	59.39	**32.74**

Bold values indicates best values found for given metric during experiments.

Within the Netronome AgilioCX 2 × 25 GbE, the primary limits are the number of microengines (60), data rate (25 GbE), and number of hardware signals (15 per thread). While the constraint on microengines can be mitigated by assigning multiple features to a microengine, the memory (i.e., number of transfer registers) and asynchronous I/O (i.e., number of hardware signals) constraints limit the depth of the trained decision tree to 480 leaf nodes. Our current implementation is limited to decision trees with 256 leaf nodes, since more complex firmware is required to assign a hardware signal to a greater number of transfer registers.

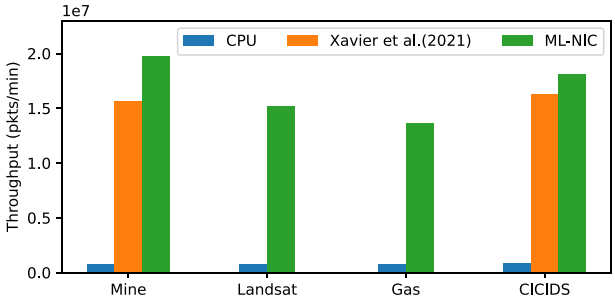

FIGURE 5
This figure provides a bar graph of the throughputs observed with our approach and two baselines.

6.5 Offloading more models

6.5.1 Decision tree

In addition to the decision tree models we have deployed in this work, we provide some more insights on other decision trees that our current work can offload onto a SmartNIC via an ablation study. In our ablation study, we use the CICIDS datasets to construct 10 decision trees with a constraint on the maximum number of leaf nodes between 16 and 256, based on the leaf node limit we mentioned in Section 3. For each decision tree, we look at the depth, number of nodes, number of leaf nodes, size of the pickle file, size of the firmware file, and SmartNIC RAM usage. We present our findings in Figure 6. Note that we scaled some of the measurements by a factor of 10, so the trends in the some of the decision tree parameters would be more clear.

From Figure 6, we observe the following. First, we note the slow inclination, followed by a brief declination, then continued inclination in the usage of SmartNIC RAM. We also observe a similar pattern of inclination, declination, then inclination again in the trend for the model firmware size. We attribute this to how we compiled two instances of the model 1 and 2 per island to maximize our usage of the computation resources on the SmartNIC. For the remaining models, we only compiled one instance per island. Since models 1 and 2 have two instances compiled per island, more RAM and instruction memory would be needed to store the labels for the leaf nodes and decision tree logic. Based on the rate of inclination between firmware size and SmartNIC RAM usage, we see that a primary concern for offloading larger models is the amount of instruction memory available per microengine. After model 2, we see that the (scaled) trend for firmware size grows slower than that of the size of model pickle file and number of nodes by 1.64 and 1.53 and grows faster than the trend for decision tree depth by 4.07 and number of leaf nodes by 1.30.

TABLE 6 Hardware utilization measurements.

Dataset	Measure	CPU	Xavier et al. (2021)	ML-NIC
Mine	Min CPU (%)	**0.25**	**0.25**	**0.25**
	Avg CPU (%)	3.82	0.77	**0.52**
	Max CPU (%)	7.56	7.75	**0.94**
	Min Host RAM (MB)	2,226.76	2,147.76	**1,908.97**
	Avg Host RAM (MB)	2,228.43	2,150.75	**1,910.71**
	Max Host RAM (MB)	2,230.56	2,153.44	**1,911.81**
	SmartNIC RAM (MB)	96.15	973.00	**21.13**
Landsat	Min CPU (%)	**0.25**	N/A	**0.25**
	Avg CPU (%)	3.81	N/A	**0.51**
	Max CPU (%)	7.53	N/A	**0.88**
	Min Host RAM (MB)	2,226.50	N/A	**1,909.86**
	Avg Host RAM (MB)	2,229.00	N/A	**1,910.69**
	Max Host RAM (MB)	2,231.77	N/A	**1,911.70**
	SmartNIC RAM (MB)	96.15	N/A	**21.12**
Gas	Min CPU (%)	0.25	N/A	**0.19**
	Avg CPU (%)	3.78	N/A	**0.51**
	Max CPU (%)	7.46	N/A	**0.88**
	Min Host RAM (MB)	2,223.59	N/A	**1,908.86**
	Avg Host RAM (MB)	2,226.08	N/A	**1,909.83**
	Max Host RAM (MB)	2,228.07	N/A	**1,910.98**
	SmartNIC RAM (MB)	96.15	N/A	**21.122**
CICIDS	Min CPU (%)	**0.19**	0.25	0.25
	Avg CPU (%)	3.68	0.80	**0.52**
	Max CPU (%)	7.20	9.09	**0.81**
	Min Host RAM (MB)	2,225.02	2,077.43	**1,907.22**
	Avg Host RAM (MB)	2,227.38	2,078.88	**1,908.36**
	Max Host RAM (MB)	2,230.01	2,080.04	**1,909.21**
	SmartNIC RAM (MB)	96.15	973.00	**21.12**

Bold values indicates best values found for given metric during experiments.

6.5.2 Other machine learning models

Besides decision trees, we also consider approaches for executing inference with other machine learning models. Since inference for many popular machine learning models relies heavily on the matrix-vector multiplication operation, we look into techniques for efficient and effective matrix-vector multiplication that can be performed by a SmartNIC. In addition to conducting inference on models such as neural network and support vector machines for supervised tasks, we find an implementing a suitable matrix-vector multiplication method necessary for unsupervised learning, such as with implementing k-means using cosine similarity as the similarity measure instead of Euclidean distance. First, a naive approach that we consider is creating a lookup table per weight to match a feature value with the multiplication of that feature with the specific weight. In this approach, the feature computation microengines would be responsible for doing the multiplication lookup based on the weight and feature value, and the result aggregation microengines would sum up the multiplied weight-feature values to obtain the final result. However, this approach may not be feasible for problems that require a large number of weights due to memory constraints on the SmartNIC.

An alternative approach would be to consider using natural logarithm and the exponent function (i.e., e^x). Instead of storing a lookup table per feature, two lookup tables can be stored to approximately compute the natural logarithm and exponent of the feature values based on their fixed range (i.e., we assume each feature is in the range $[0, 1]$ in Section 3). Then, each feature computation microengine would be operate on a specific feature by conducting a lookup for the natural logarithm of the feature, taking the sum of the natural logarithms of the weights and the feature value, and conduct a lookup of the exponent of the sum of the natural logarithm values. While this approach may resolve issues with the memory constraint, a large number of features requires multiple lookups to the memory region holding the tables (i.e., CLS or IMEM) that can congest the CPP bus. So, to avoid this issue, the lookup tables could be replaced with first-order taylor approximations of the natural logarithm and exponent functions for a specific number of reference points in the range $[0, 1]$, where the taylor approximations can be represented using additions and bit shifts. At the same time though, the use of first-order taylor approximations can result in more erroneous model predictions.

More recently though, work by Blalock and Guttag (2021) proposed a novel technique for matrix-matrix multiplication that used locality-sensitive hashing to determine suitable functions [denote as $g(A)$] that can be executed efficiently using balanced binary regression trees. Based on their findings and our current implementation for decision tree inference, we believe their approach to be a more promising direction for executing matrix-vector multiplication on a SmartNIC.

6.6 Productionization and scaling

When deploying decision tree models in the real world, we consider factors such as model updates and scaling. Concerning model updates (i.e., models retrained on larger datasets), we still limit the number of leaf nodes to 256, which may or may not be helpful as a regularization technique to prevent decision tree overfitting. In order to deploy a new decision tree, the original decision tree (i.e., the model firmware file) needs to be unloaded from the SmartNIC, and then the firmware for the new decision tree can be loaded onto the SmartNIC. This means that SmartNIC would be inactive while unloading the old decision tree and loading the new decision tree. So, inference requests can not be handled by a SmartNIC during that time.

For scaling, we primarily focus on the first model deployment scenario (model fits on an island). We do not believe much scaling of SmartNIC resources can be done as the entire card is required for one instantiation of the model. In the first deployment scenario,

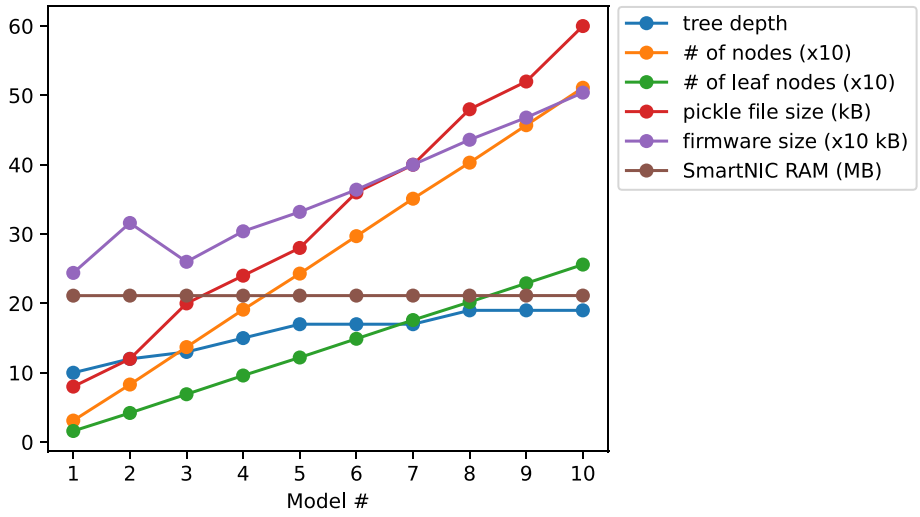

FIGURE 6
This figures shows how ML-NIC scales with respect to different decision tree parameters.

though, based on the amount of network traffic, firmware can be developed to set aside some islands for the decision tree model and others for other tasks. However, like the model update case, old firmware would need to be unloaded and new firmware loaded. So, inference requests can not be handled by a SmartNIC while unloading old and loading new firmware.

6.7 Hardware improvements

Considering SmartNICs with Micro-C support that possess additional hardware capabilities, we first note SmartNICs with additional programmable flow-processing microengines. A SmartNIC can include more programmable microengines in two ways: additional islands or microengines per island. With additional microengines per island, we expect a further reduction in latency for all three model deployment scenarios. This would happen because more models would be able to fit on an island, which would remove the need for communication between the microengines on different islands. Communication between microengines on different islands is more expensive than between microengines on the same island. With additional islands, we expect a further increase in throughput for model deployment scenario one (the model can fit on one island). More islands mean more instances of the model that can be instantiated on the SmartNIC, which would allow it to meet more inference requests. In either scenario, we expect larger trees (with respect to the number of features) to be more easily deployed, given that each microengine would be responsible for analyzing fewer features.

Next, we consider a SmartNIC with additional transfer registers per microengine. More transfer registers means that fewer microengines would be needed to perform result aggregation based on the feature analysis conducted by the feature computation microengines. With greater availability of microengines, we could likely deploy larger trees on the SmartNIC.

Lastly, another hardware improvement we consider is the addition of one or more FPUs. Adding FPUs resolves the issue with minor effectiveness degradation that we observe with the current iteration of our work while likely maintaining similar latency and throughput performance that we've observed in this work.

7 Related work

Recently, several works have leveraged programmable data planes to make aspects of machine learning more efficient. These works can be split into two categories: model training and model inference. We focus on the latter. Within model inference, research efforts are focused on leveraging how entire or portions of the machine learning model inference process can be offloaded onto programmable data planes while maintaining adequate model performance. These implementations are most commonly conducted on programmable switches and SmartNICs.

7.1 Programmable switch

For works that map machine learning models onto programmable switches, we generally observe a focus on specific models used to address certain tasks. We first note Net2Net (Siracusano and Bifulco, 2018) that proposed quantizing neural networks into binary neural networks, since they require operations that are readily available on modern switching chips. Rather than quantizing a trained neural network into a binary neural network, Qin et al. (2020) directly trained binary neural networks and mapped them onto the data planes of programmable switches using P4 to handle the network intrusion detection use case. As an alternative to binary neural network quantization, Dao et al. (2021) used neuron pruning to map neural networks onto programmable switches for the network intrusion use case. While the above neural network works solely considered a single

programmable switch for deployment, Saquetti et al. (2021) implemented a neural network neuron distribution method across multiple programmable switches and coordinated inference of the model between the switches to optimize resource usage. Similarly, JointNIDS (Dao and Lee, 2022) also employed a distributed neural network inference approach. But, the neural network intrusion detection models were split into two sequential sub-models with overlapping hidden units and mapped the sub-models onto two programmable switches. The authors assigned one programmable switch to detect major network attacks, while the second handled the more subtle aspects of network traffic classification. Rather than staying within the constraints of off-the-shelf programmable network devices, Taurus (Swamy et al., 2022) extended the PISA architecture of programmable switches by adding custom hardware to support parallelism and additional operations (i.e., multiplication, nonlinear operations) needed to run neural network inference without any quantization.

Regarding tree-based models, pForest (Busse-Grawitz et al., 2019) developed a optimization technique to map a random forest classifier to a programmable switch in P4 for network flow classification. Furthermore, this approach adaptively switches out the current classifier with others based on the network flows observed. In addition to mapping a random forest classifier, Planter (Zheng and Zilberman, 2021) mapped a xgboost and isolation forest classifier to programmable switches using overlapping trees to overcome some of the inefficiencies observed in pForest (Busse-Grawitz et al., 2019). Similar to pForest (Busse-Grawitz et al., 2019) and SMASH (Kamath and Sivalingam, 2021) also focused on the network flow classification task and used an improved hash-and-store algorithm with a decision tree model for early flow classification. Also working with decision trees, pHeavy (Zhang et al., 2021) implemented trained decision trees on the data plane to reduce the overhead involved with communicating to the control plane in Software-Defined Networking (SDN) when classifying highly-congested network flows. In contrast to the other tree-based model offloading approaches that focus on network-related use cases, NetPixel (Siddique et al., 2021) implemented decision trees on P4 programmable switches to handle image classification. To address some of the issues with deployment of decision trees and other machine learning algorithms onto programmable data planes, Mousika (Xie et al., 2022) introduced a teacher-student knowledge distillation approach to translate machine learning models to binary decision trees, which are more suitable for mapping onto the data plane.

On top of supervised machine learning, the deployment of unsupervised learning algorithms onto programmable switches has also been explored. Clustreams (Friedman et al., 2021) used a combination of the quadtree data structure and a match+action table stored in Ternary Content Addressable Memory (TCAM) to cluster network traffic efficiently. In addition, ACC-Turbo (Alcoz et al., 2022) redesigned the original Aggregate-based Congestion Control (ACC) approach using online clustering and a scheduling algorithm to mitigate pulse-wave DDoS attacks.

Unlike the works above that focus on a specific machine learning algorithm type (i.e., neural networks, tree-based models, clustering, etc.), IIsy (Xiong and Zilberman, 2019) introduced mapping schemes for several machine learning algorithms, such as decision trees, k-means, naive bayes, and support vector machines, to the data plane using the match-action pipeline in programmable switches. Also, Hong et al. (2024) developed a feature engineering and model deployment strategy for tree-based models (i.e., decision trees, random forests, xgboost), k-nearest neighbor, and k-means to handle the high-frequency stock market trading task.

7.2 SmartNIC

Similar to the works that address deployment of machine learning model inference onto programmable switches, we see works about model inference onto SmartNICs that also consider neural networks and decision trees used for particular applications. Using the approach proposed by Net2Net (Siracusano and Bifulco, 2018), BaNaNa split (Sanvito et al., 2018) accelerated the inference of neural networks by splitting a neural network at its fully-connected layers, sending all prior layers to the host system's CPU for inference, and quantizing the fully-connected layers to run portion of the inference on the host system's SmartNIC. Different from the other works that primarily look into P4 implementations, N3IC (Siracusano et al., 2022) used Micro-C and P4 to map binary neural networks onto a greater variety of targets (i.e., SmartNICs) for traffic analysis use cases. Regarding tree-based models, Xavier et al. (2021) presented a framework for deploying decision tree models onto SmartNICs in P4. The authors demonstrated that their framework can achieve high accuracy (above 95%) in a network intrusion detection use case. While similar to our work, we note that ML-NIC works on a greater variety of use cases outside of network traffic analysis. Furthermore, ML-NIC's Micro-C implementations can parallelize the model inference process, which is not possible with P4. While the works on machine learning inference offloading for SmartNICs do not cover unsupervised learning to our knowledge, they do address traditional reinforcement learning. Opal (Simpson and Pezaros, 2022) implemented online reinforcement learning onto a SmartNIC data plane, relying on classical reinforcement algorithms such as Sarsa (Sutton, 2018) and avoiding neural networks.

8 Future directions

8.1 Implementing additional models

As mentioned in Section 6, our next step is to expand our framework to other machine learning algorithms, such as support vector machines and neural networks. We find that implementing the approximate matrix-matrix multiplication approach developed by Blalock and Guttag (2021) to be means of achieving this goal. In addition to matrix-matrix multiplication, there are floating point operations that are often performed for various models such as neural networks. Thus, in order to implement such models, future work may involve, either adding hardware support for these operations or using quantized operations as a default. We discuss the potential future work in this area in Section 8.2.

8.2 Improving the floating-point representation

While our current approach leverages fixed-point 16-bit features and produces comparable model effectiveness scores to the baselines, it is possible that the proposed float-representation scheme can be improved without adding more hardware. One potential future direction to be explored is to look into the posit representation (Gustafson and Yonemoto, 2017) as a potential alternative, since it resolves the issue of NaN quantities observed in the standard floating-point representation. Also, while implementing the standard floating-point representation on the SmartNIC may seem like a viable solutions, we believe that the float-pointing representation standard would introduce additional latency due to the additional computation spent managing mantissa bits, exponent bits, and NaN quantities. In addition, adding any additional hardware support may cause added cost and energy consumption of SmartNIC, which defeats the purpose of using the SmartNIC in the first place. Thus, a software/algorithmic based approaches for performing floating point operations will be a great direction for future work.

8.3 Automating the model deployment

Furthermore, despite automating the process of decision tree logic extraction, the process of building the model mapping still mostly requires the developer to manually allocate cores as one of packet collection, feature computation, or result aggregation. We think the model mapping component can be made more efficient with additional code that considers the computation constraints of the SmartNIC and presents a mapping scheme to remove some of the tedious work in deploying a model onto the SmartNIC. Thus, a direction for future work may involve building a more sophisticated system that can perform model compilation, optimization and deployment automatically to the SmartNIC. This work can be further strengthened by adding a notion of distributed deployment and model inference across multiple SmartNICs located on multiple server.

8.4 Utilizing different types of SmartNICs

While this work primarily utilizes ASIC-based SmartNIC, it is possible to implement similar work on other ASIC-based and other types of SmartNICs, such as FPGA-based SmartNICs. While the optimizations we performed in this paper is specific to Netronome SmartNIC, the overall idea of mapping memory into different SmartNIC region is quite generic. Thus, a potential valuable future work is to perform similar optimization strategies across different types of hardware implementations to understand the similarities and differences in the effectiveness of the proposed optimization and compilation strategies.

9 Conclusion

Low-latency model inference is a necessity for many time-sensitive machine learning applications. This paper demonstrates that ML-NIC is a suitable framework for performing machine learning model inference. Our evaluation of the first iteration of ML-NIC shows that it can deploy larger models than the state-of-the-art SmartNIC approach, can produce predictions at faster speeds with a minor loss in model effectiveness compared to the CPU solution, and is robust to high network utilization and CPU loads.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

RK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. DA: Investigation, Supervision, Writing – original draft, Writing – review & editing. SC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project has been funded by the NSF CRII 2245352.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1493399/full#supplementary-material

References

Alcoz, A. G., Strohmeier, M., Lenders, V., and Vanbever, L. (2022). "Aggregate-based congestion control for pulse-wave ddos defense," in *Proceedings of the ACM SIGCOMM 2022 Conference*, 693–706. doi: 10.1145/3544216.3544263

Blalock, D., and Guttag, J. (2021). "Multiplying matrices without multiplying," in *International Conference on Machine Learning* (PMLR), 992–1004.

Breiman, L., Jerome Friedman, C. J. S., and Olshen, R. (1984). *Classification and Regression Trees*. New York: Chapman and Hall/CRC.

Busse-Grawitz, C., Meier, R., Dietmüller, A., Bühler, T., and Vanbever, L. (2019). pforest: In-network inference with random forests. *arXiv preprint arXiv:1909.05680*.

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi: 10.1145/2939672.2939785

Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., et al. (2014). "Dadiannao: a machine-learning supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture* (IEEE), 609–622. doi: 10.1109/MICRO.2014.58

Choquette, J., Giroux, O., and Foley, D. (2018). Volta: performance and programmability. *IEEE Micro* 38, 42–52. doi: 10.1109/MM.2018.022071134

Corigine (2020). *Corigine nfp-4000 flow processor*. Technical report, Corigine.

Dang, H. T., Bressana, P., Wang, H., Lee, K. S., Zilberman, N., Weatherspoon, H., et al. (2020). P4xos: consensus as a network service. *IEEE/ACM Trans. Network.* 28, 1726–1738. doi: 10.1109/TNET.2020.2992106

Dao, T.-N., Hoang, V.-P., Ta, C. H. (2021). "Development of lightweight and accurate intrusion detection on programmable data plane," in *2021 International Conference on Advanced Technologies for Communications (ATC)* (IEEE), 99–103. doi: 10.1109/ATC52653.2021.9598239

Dao, T.-N., and Lee, H. (2022). Jointnids: efficient joint traffic management for on-device network intrusion detection. *IEEE Trans. Vehic. Technol.* 71, 13254–13265. doi: 10.1109/TVT.2022.3198266

Fowers, J., Ovtcharov, K., Papamichael, M., Massengill, T., Liu, M., Lo, D., et al. (2018). "A configurable cloud-scale DNN processor for real-time ai," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (IEEE), 1–14. doi: 10.1109/ISCA.2018.00012

Friedman, R., Goaz, O., and Rottenstreich, O. (2021). "Clustreams: data plane clustering," in *Proceedings of the ACM SIGCOMM Symposium on SDN Research (SOSR)*, 101–107. doi: 10.1145/3482898.3483356

Gustafson, J. L., and Yonemoto, I. T. (2017). Beating floating point at its own game: posit arithmetic. *Supercomput. Front. Innov.* 4, 71–86. doi: 10.14529/jsfi170206

He, Z., Sidler, D., István, Z., and Alonso, G. (2018). "A flexible k-means operator for hybrid databases," in *2018 28th International Conference on Field Programmable Logic and Applications (FPL)* (IEEE), 368–3683. doi: 10.1109/FPL.2018.00069

Hong, X., Zheng, C., Zohren, S., and Zilberman, N. (2024). "Accelerating machine learning for trading using programmable switches," in *27th European Conference on Artificial Intelligence (ECAI 2024)* (IOS Press), 3429–3436. doi: 10.3233/FAIA240894

Jin, X., Li, X., Zhang, H., Soulé, R., Lee, J., Foster, N., et al. (2017). "Netcache: balancing key-value stores with fast in-network caching," in *Proceedings of the 26th Symposium on Operating Systems Principles*, 121–136. doi: 10.1145/3132747.3132764

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., et al. (2017). "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*, 1–12.

Kamath, R., and Sivalingam, K. M. (2021). "Machine learning based flow classification in dCNS using p4 switches," in *2021 International Conference on Computer Communications and Networks (ICCCN)* (IEEE), 1–10. doi: 10.1109/ICCCN52240.2021.9522272

Kim, C., Sivaraman, A., Katta, N., Bas, A., Dixit, A., Wobker, L. J., et al. (2015). "In-band network telemetry via programmable dataplanes," in *ACM SIGCOMM*, 1–2. doi: 10.1145/2774993.2775007

Langlet, J. (2019). *Towards machine learning inference in the data plane* (Student thesis). Karlstad University, Department of Mathematics and Computer Science (from 2013).

Liu, Y. H. (2017). *Python Machine Learning by Example*. Birmingham: Packt Publishing Ltd.

Netronome Systems, I. (2024). *Agilio cx 2x25gbe smartnic*. Technical report, Netronome Systems, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qin, Q., Poularakis, K., Leung, K. K., and Tassiulas, L. (2020). "Line-speed and scalable intrusion detection at the network edge via federated learning," in *2020 IFIP Networking Conference (Networking)* (IEEE), 352–360.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi: 10.1007/BF00116251

Rodríguez-Luján, I., Fonollosa, J., Vergara, A., Homer, M. L., and Huerta, R. (2014). On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. *Chemometr. Intellig. Lab. Syst.* 130, 123–134. doi: 10.1016/j.chemolab.2013.10.012

Roth, D. (2016). *Decision trees*.

Sanvito, D., Siracusano, G., and Bifulco, R. (2018). "Can the network be the ai accelerator?" in *Proceedings of the 2018 Morning Workshop on In-Network Computing*, 20–25. doi: 10.1145/3229591.3229594

Saquetti, M., Canofre, R., Lorenzon, A. F., Rossi, F. D., Azambuja, J. R., Cordeiro, W., et al. (2021). Toward in-network intelligence: running distributed artificial neural networks in the data plane. *IEEE Commun. Lett.* 25, 3551–3555. doi: 10.1109/LCOMM.2021.3108940

Sharafaldin, I., Lashkari, A. H., Ghorbani, A. A., et al. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* 1, 108–116. doi: 10.5220/0006639801080116

Siddique, H., Neves, M., Kuzniar, C., and Haque, I. (2021). "Towards network-accelerated ml-based distributed computer vision systems," in *2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS)* (IEEE), 122–129. doi: 10.1109/ICPADS53394.2021.00021

Simpson, K. A., and Pezaros, D. P. (2022). "Revisiting the classics: online RL in the programmable dataplane," in *NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium* (IEEE), 1–10. doi: 10.1109/NOMS54207.2022.9789930

Siracusano, G., and Bifulco, R. (2018). In-network neural networks. *arXiv preprint arXiv:1801.05731*.

Siracusano, G., Galea, S., Sanvito, D., Malekzadeh, M., Antichi, G., Costa, P., et al. (2022). "Re-architecting traffic analysis with neural network interface cards," in *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 513–533.

Srinivasan, A. (1993). Statlog (Landsat Satellite). *UCI Machine Learning Repository*.

Sutton, R. S. (2018). *Reinforcement Learning: An Introduction*. Menomonie, WI: A Bradford Book.

Swamy, T., Rucker, A., Shahbaz, M., Gaur, I., and Olukotun, K. (2022). "Taurus: a data plane architecture for per-packet ml," in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 1099–1114. doi: 10.1145/3503222.3507726

Tong, D., Qu, Y. R., and Prasanna, V. K. (2017). Accelerating decision tree based traffic classification on fPGA and multicore platforms. *IEEE Trans. Parallel Distr. Syst.* 28, 3046–3059. doi: 10.1109/TPDS.2017.2714661

Wray, S. (2014). *The joy of micro-c*. Technical report, Netronome Systems, Inc.

Xavier, B. M., Guimarães, R. S., Comarela, G., and Martinello, M. (2021). "Programmable switches for in-networking classification," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications* (IEEE), 1–10. doi: 10.1109/INFOCOM42981.2021.9488840

Xie, G., Li, Q., Dong, Y., Duan, G., Jiang, Y., and Duan, J. (2022). "Mousika: enable general in-network intelligence in programmable switches by knowledge distillation," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, 1938–1947. IEEE. doi: 10.1109/INFOCOM48880.2022.9796936

Xiong, Z., and Zilberman, N. (2019). "Do switches dream of machine learning? Toward in-network classification," in *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, 25–33. doi: 10.1145/3365609.3365864

Yilmaz, C., Kahraman, H. T., and Söyler, S. (2018). Passive mine detection and classification method based on hybrid model. *IEEE Access* 6, 47870–47888. doi: 10.1109/ACCESS.2018.2866538

Zhang, B., Kannan, R., Prasanna, V., and Busart, C. (2023). "Accelerating gnn-based sar automatic target recognition on hbm-enabled fpga," in *2023 IEEE High Performance Extreme Computing Conference (HPEC)* (IEEE), 1–7. doi: 10.1109/HPEC58863.2023.10363615

Zhang, X., Cui, L., Tso, F. P., and Jia, W. (2021). pheavy: predicting heavy flows in the programmable data plane. *IEEE Trans. Netw. Serv. Manag.* 18, 4353–4364. doi: 10.1109/TNSM.2021.3094514

Zheng, C., and Zilberman, N. (2021). "Planter: seeding trees within switches," in *Proceedings of the SIGCOMM '21 Poster and Demo Sessions, SIGCOMM '21* (New York, NY, USA: Association for Computing Machinery), 12–14. doi: 10.1145/3472716.3472846