**Literature Review**

**IoT**

The IoT is a system of smart physical devices connected to the internet worldwide, all gathering and exchanging information on a large scale. As a result of this vast paradigm, more applications and services of IoT are emerging.

Tama and Rhee (2019) [14] presented an anomaly-based IDS that, as a detection engine, use a gradient boosted machine (GBM). Grid search was used to get the optimal GBM parameters. The presented IDS performance is validated on 20% of three separate datasets: NSL-KDD, GPRS, and UNSW-NB15, using hold-out and tenfold cross-validation. The authors also show that the presented IDS surpasses the precision, specificity, sensitivity, false alarm rate and area under the curve metric of the fuzzy classifier, tree-based ensemble classifier, and GAR forest with a detection accuracy of 91.82% in the KDDTest+, 86.51% in KDDTest-21 dataset, 91.31% in UNSW\_NB15\_Test dataset, and 82.6% in the GPRS dataset.

Moustafa et al. (2018) [15] presented an ensemble IDS that alleviates mischievous events like botnet attacks in Hypertext Transfer Protocol, Message Queue Telemetry Transport, and Domain Name System protocols that use backend database structures of IoT infrastructure to accumulate data generated from IoT devices. Using DT, NB, and ANN ML algorithms, they have developed an ensemble model that evaluates the learned statistical flow features from the aforementioned protocols as well as detecting mischievous events with a detection accuracy of 99.54% in DNS protocol based feature and 98.97% detection accuracy in HTTP protocol based feature using UNSW-NB15 dataset. They have also achieved 98.29% detection accuracy in the DNS protocol based feature and 98.36% detection accuracy on HTTP data source of NIMS dataset.

Hasan et al. (2019) [16] compared the performance of ANN, SVM, LR, DT, and RF ML algorithms in detecting attack & anomaly in IoT sensors for multi-class attacks: denial of service, malicious control, scan, data type probing, malicious operation, spying, and wrong setup. They have used a synthetic dataset prepared by Pahl et al. (2018) [17] in a computer-generated IoT background using Distributed Smart Space Orchestration System, where communication among microservices occurs using Message Queue Telemetry Transport protocol. Logistic regression (LR) achieved the highest accuracy with an anomaly detection rate of 98.3%.

Alrashdi et al. (2019) [18] proposed an anomaly recognition approach to detect vulnerable IoT devices in smart cities that monitor IoT traffic at dispersed fog nodes using the Random Forest (RF) ML algorithm. Between the cloud and loT layers, fog nodes work to reduce energy usage, capacity, and bandwidth. It is a binary label classification problem with only two labels: normal and attack. RF model achieved an F1 score of 0.98 using the UNSW-NB15 dataset.

A novel 2-layer dimension reduction & 2-tier detection module based IDS is presented by Pajouh et al. (2016) [19]. It can identify mischievous events such as Remote to Local & User to Root attacks. In this model, the high dimensional dataset is converted into a lower dimension with reduced features by linear discriminate and component analysis. Naïve Bayes and KNN with confidence factor that provides a number in the range of -1 to 1, were used to detect mischievous events using the NSL-KDD dataset. At first-tier, Naive Bayes is used to detecting anomalies with the linear discriminant analysis applied to it to reduce features. At the second tier, KNN is used for the classification from that lesser featured data. The model achieved a detection accuracy of 70.15% in User to Root attack and 42% in Remote to Local attack.

In the next part of the paper, we present a comparative analysis of our reviewed works.

**Comparative Analysis**

TABLE I: Objective & Data Set Properties (IoT)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Objective** | **ML Method** | **Training dataset** | **Feature Selection from Dataset** | **Outcome** | **Reference** |
| Implement a competent IDS based on Gradient Boosted Machine (GBM). | GBM | NSL-KDD, GPRS, and UNSW-NB15. | Not mentioned. | Binary classification: Normal and Attack. | [14] |
| Develop an ensemble IDS that lessens mischievous events in IoT infrastructure. | Ensemble method. | UNSW-NB15 | Statistical flow features using Correlation Coefficient such as MQTT and service based DNS and HTTP feature set. | Multiclass classification. | [15] |
| Performance comparison of ML methods in IoT Sensor network. | ANN, SVM, LR, DT, and RF. | DS2OS | 13 features.  12 – Object type  1- int64 type | Multiclass classification. | [16] |
| Detection of vulnerable IoT devices in smart IoT environment. | RF | UNSW-NB15 | Nominal, Integer, Binary and Float. | Binary classification: Normal and Attack. | [18] |
| Develop a 2-layer dimension reduction & 2-tier detection module to reduce feature and intrusion detection from Remote to Local & User to Root attacks. | NV, KNN with Confidence Factor. | NSL-KDD | Principal Component Analysis was used to extract features. Then by linear discriminant analysis was used to reduce those features into {lda1, lda2, lda3, lda4}. | Multiclass classification:  Normal, DoS, Probe, U2R and R2L. | [19] |

TABLE II: Performance Comparison of ML method (IoT)

|  |  |  |  |
| --- | --- | --- | --- |
| **ML Method**  **Category** | **ML Method** | **Classification accuracy** | **Reference** |
| Supervised | GBM | 97.88% | [14] |
| RF | 97.18% |
| DNN | 97.48% |
| CART | 97.38% |
| SVM | 91.34% |
| Supervised | Ensemble | 99.64% [UNSW-NB15-DNS] | [15] |
| 98.27% [UNSW-NB15-HTTP] |
| 98.29% [NIMS-DNS] |
| 98.36% [NIMS-HTTP] |
| Supervised | DT | 99.40% | [16] |
| SVM | 98.20% |
| LR | 98.30% |
| ANN | 99.40% |
| RF | 99.41% |
| Supervised | RF | 98.00% | [18] |
| Supervised | Combination of NV &  KNN | 70.15% in User to Root attack | [19] |
| 42% in Remote to Local attack. |