## Appendix:

## A.1 Data Extraction form

Table 1. Data type and item extracted from each study

| Data Type | ID | Data Item | Description |
|---|---|---|---|
| Context | D1-D7 | Title, author, venue, publication year, publisher, summary, open challenges | Title, author, venue, publication year, publisher, summary including aim, strength, and weakness of the study and Open challenges to be resolved in future. |
| RQ1 | D8-D9 | Features, Feature Engineering Method | Features and feature engineering method: automatic or manual used to implement NLP-based HIDS |
| | D10-D12 | Learning Type, Classifier Type, Detection technique | Type of Learning method, Classifier type (e.g., Base, Ensemble), detection technique used for intrusion detection |
| | D13-D14 | HIDS type, Attack detection/classification | Type of HIDS (misuse, anomaly), attack detection (e.g., benign, malicious) or classification (detect specific attack) |
| RQ2 | D15 | Attacks | Attacks that are targeted to be detected |
| RQ3 | D16-D17 | Data Source, Dataset | Data source or dataset used for training or testing HIDS |
| RQ4 | D18 | Evaluation Metric | Metrics used for evaluating HIDS |

## A.2 Feature types with mapped studies

Table 2. Feature types used in NLP-based HIDS with mapped studies

| Feature Type | Study Ref |
|---|---|
| Statistical (22) | S1, S2, S3, S6, S7, S14, S20, S27, S32, S33, S36, S37, S40, S41, S42, S49, S52, S65, S80, S81, S86, S87 |
| Contextual (49) | S4, S5, S8, S9, S10, S11, S13, S15, S16, S17, S19, S21, S22, S24, S25, S26, S28, S29, S34, S35, S39, S43, S44, S45, S47, S50, S54, S58, S59, S60, S62, S63, S64, S67, S68, S70, S71, S72, S73, S74, S76, S79, S82, S84, S88, S93, S94, S95, S97 |
| Attribute (1) | S18 |
| Temporal (3) | S12, S55, S66 |
| Statistical+Contextual (7) | S46, S61, S77, S78, S85, S96, S98, |
| Statistical+Attribute (4) | S30, S48, S53, S99 |
| Statistical+Attribute+Temporal (3) | S57, S75, S91 |
| Contextual+Attribute (4) | S23, S31, S51, S56 |
| Temporal+Contextual (3) | S38, S69, S83 |
| Statistical+Contextual+Attribute+Temporal (3) | S89, S90, S92 |

### A.3 Detection techniques categories, sub-categories, instances mapped with reviewed studies

Table 3. Detection techniques categories, sub-categories, and instances mapped with the reviewed studies

| Detection Category | Sub-category | Technique Instances | Study Ref |
|---|---|---|---|
| Traditional ML (67)<br>S1, S2, S3, S5, S6, S7, S11, S14, S18, S19, S21, S22, S23, S24, S25, S26, S27, S28, S30, S31, S32, S33, S34, S35, S36, S37, S39, S40, S41, S42, S44, S45, S46, S48, S52, S53, S57, S59, S60, S61, S62, S63, S64, S65, S69, S70, S71, S72, S75, S77, S78, S79, S84, S85, S86, S87, S88, S89, S90, S91, S92, S93, S94, S95, S96, S97, S99 | Bayesian (16)<br>S3, S19, S21, S23, S30, S31, S32, S45, S48, S60, S62, S65, S69, S70, S91, S99 | Naive Bayes (6) | S19, S32, S60, S62, S65, S91 |
| | | BernoulliNB | S65 |
| | | MultinomialNB | S66 |
| | | GaussianNB | S65 |
| | | GaussianProcessClassifier | S65 |
| | | Gaussian Mixture Models (GMM) | S30, S48, S99 |
| | | ComplementNB | S65 |
| | | SC2/SC2.2/Markov-Bayes 5 | S21, S23, S31, S45, S70 |
| | | time Bayesian networks (CTBN) | S69 |
| | | Multi-variable Naïve Bayesian (MNB) | S3 |
| | Rule system (10)<br>S3, S6, S7, S19, S53, S59, S60, S62, S65, S92<br>("Decision Tree (8)<br>S3, S6, S19, S53, S60, S62, S65, S92"<br>"Rule system (4)<br>S7, S59, S60, S62") | ExtraTreeClassifier | S65 |
| | | Decision Tree/C4.5/C5 | S3, S6, S19, S53, S60, S62, S65, S92 |
| | | PART | S60 |
| | | RIPPER | S60, S62 |
| | | OneR | S62 |
| | | ZeroR | S62 |
| | | Rough Set Classification (RSC) | S7, S59 |
| | Instance-based (13)<br>S2, S3, S24, S32, S33, S36, S41, S52, S62, S65, S85, S87, S99 | K-Nearest Neighbors (KNN) | S2, S3, S24, S32, S33, S36, S41, S52, S62, S65, S85, S87, S99 |
| | | K-furthest neighbors (KFN) | S85 |
| | | K-centers | S87 |
| | Ensemble (12)<br>S1, S3, S5, S21, S60, S63, S64, S65, S75, S89, S90, S94 | Random Forest (RF) | S3, S60, S64, S65, S75, S94 |
| | | Isolation Forest (IF) | S1, S5, S63 |
| | | ExtraTreesClassifier | S65 |
| | | AdaBoost | S21. S65 |

| | | | |
|---|---|---|---|
| | | Bagging Classifier | S65 |
| | | GradientBoostingClassifier | S65 |
| | | XGBoost | S65, S89, S90 |
| | Statistical model (15) S11, S22, S25, S26, S28, S35, S39, S42, S44, S46, S53, S79, S88, S95, S96 | Clustered Markov Networks (CMN) /CMN with Outlying Subspace (CMN-OS) | S42 |
| | | Markov chain | |
| | | CRF (conditional random fields) | S35 |
| | | HMM/I-HMM | S11, S22, S25, S26, S28, S39, S44, S46, S53, S79, S88, S95, S96 |
| | Clustering (12) S23, S24, S31, S36, S42, S48, S62, S72, S77, S84, S93, S99 | Harmony Search based K-means clustering | S72 |
| | | k-means (9) | S24, S36, S42, S48, S62, S77, S84, S93, S99 |
| | | distance/RE based | S23, 31 |
| | | Fuzzy clustering | S24 |
| | Support Vector Machine (24) S1, S2, S3, S5, S6, S19, S30, S32, S33, S37, S46, S48, S57, S60, S61, S62, S63, S64, S71, S85, S86, S91, S92, S96 | Support Vector Machine (SVM) (10) | S2, S3, S6, S19, S32, S33, S60, S62, S64, S71 |
| | | One Class SVM (OCSVM) (12) | S1, S5, S30, S37, S46, S48, S61, S63, S85, S91, S92, S96 |
| | | SMO (Sequential Minimal Optimization) | S62 |
| | | SVDD | S57, S86 |
| | NN (9) S3, S6, S14, S19, S27, S32, S34, S40, S97 | Multi-layer Perceptron (MLP)/ NN 8 | S3, S14, S19, S6, S32, S34 |
| | | extreme learning machine (ELM) 4 | S27, S32, S40, S97 |
| | Miscellaneous (6) S1, S2, S18, S42, S65, S87 | Logistic reg | S2, S65 |
| | | axis aligned bounding box | S87 |
| | | Optimization Algorithm Based on Bee Stinging (OABBS) | S18 |
| | | Label Propagation Algorithm (LPA) | S42 |

| | | | |
|---|---|---|---|
| | | Calibrated Classifier CV | S65 |
| | | Linear Discriminant Analysis | S65 |
| | | Quadratic Discriminant Analysis | S65 |
| | | Local Outlier Factor (LOF) | S1 |
| DL (22)<br>S4, S5, S8, S10, S11, S15, S17, S58, S63 , S64, S65, S67, S68, S71, S73, S74, S78, S80, S82, S93, S94, S98 | seq2seq Language modeling (5)<br>S5, S63, S64, S73, S94 | RNN/ RNN-VED /LSTM/GRU /BiLSTM/CuDNNLSTM 5 | S5 , S63, S64, S73, S94 |
| | Deep Neural network (18)<br>S4, S8, S10, S11 , S17, S58, S64, S65, S68, S71, S74, S78, S80, S82, S93, S94, S98 | Deep Multi-layer Perceptron/DNN | S65, S78, S98 |
| | | CNN/FCN/TCN(temporal convolutional neural network) (8) | S4, S8, S10, S17, S64, S71, S82, S94 |
| | | RNN/LSTM/GRU (8) | S4, S10, S11, S58, S64, S68, S74, S93 |
| | | Autoencoder /Variational Autoencoder (2) | S58, S93 |
| | | DBN (Deep belief network) | S80 |
| | Ensemble NN (3)<br>S4, S67, S15 | LSTM-FCN, GRU-FCN | S4 |
| | | CNN-LSTM, CNN-GRU, CNN-LSTM-NN | S4, S67, S15 |
| Rule-based (27)<br>S9, S12, S13, S16, S20, S25, S26, S29, S38, S43, S46, S47, S49, S50, S51, S54, S55, S56, S59, S66, S76, S77, S81, S83, S85, S96, S98 | Semantic Ontology (4) | | S13, S29, S43, S55 |
| | Model/language-based (11) | | S26, S38, S50, S77, S9, S20, S47, S51, S54, S59, S81 |
| | Sequence based (12) | | S12, S16, S25, S46, S56, S96, S98, S26, S76, S85, S49, S66, S83 |

### A.4 Data source, availability, dataset type, and instance mapped with reviewed studies

Table 4. Data source, availability, dataset type, and instance mapped with reviewed studies

| Data Source | Availability | Dataset Type | Instance | Study Ref |
|---|---|---|---|---|
| Sys call | public | Real | AWSCTD | S4, S10 |
| | | Sim | UNM | S12, S16, S21, S22, S24, S25, S34, S35, S38, S39, S40, S45, S51, S54, S70, S72, S76, S77, S81, S83, S84, S86, S88, S98 |
| | | | Firefox DS | S25, S38 |
| | | | FIT-UTK | S51 |
| | | Hyb | ADFA-LD | S1, S2, S3, S5, S6, S7, S8, S11, S14, S15, S16, S17, S19, S25, S26, S27, S32, S33, S36, S37, S40, S41, S46, S52, S58, S60, S61, S62, S63, S64, S65, S67, S68, S71, S73, S74, S78, S79, S80, S82, S85, S87, S89, S90, S92, S93, S94, S95, S98 |
| | | | ADFA-WD | S1, S6, S21, S62, S78, S92 |
| | | | CANALI-WD | S46, S79, S96 |
| | private | Real | Customized | S42 |
| | | Sim | Customized | S22, S26, S28, S39, S44, S49, S50, S58, S59 |
| Audit | public | Real | Vergina | S30, S48, S99 |
| | | | thmmy | S30, S99 |
| | | | www_ee | S30, S99 |
| | | | PUS | S83 |
| | | Sim | DARPA | S23, S31, S32, S40, S51, S57, S69 |
| | | | NGIDS-DS | S1, S17 |
| | private | Real | Customized | S13, S29, S53 |
| | | Sim | Customized | S9, S47, S55, S56, S57 |
| | | Hyb | Customized | S75 |
| Sys log | private | Real | Customized | S20, S91 |
| | | Sim | Customized | S18, S43 |

### A.5 Evaluation Metrics of intrusion detection with mapped studies

Table 5. Evaluation Metrics of intrusion detection with mapped studies

| | | |
|---|---|---|
| Detection Performance | Detection Rate (Recall, detection accuracy, TPR, true positive rate) | S2, S3, S4, S5, S6, S7, S8, S11, S12, S15, S17, S18, S19, S20, S21, S22, S23, S25, S27, S30, S31, S32, S33, S34, S35, S36, S37, S38, S39, S40, S41, S42, , S44, S45, S46, S47, S48, S52, S53, S57, S60, S61, S62, S63, S67, S68, S70, S71, S72, S73, S74, S76, S77, S79, S80, S81, S82, S85, S86, S87, S89, S90, S91, S92, S93, S98, S99 |
| | False Alarm Rate (FAR, FPR, false positive rate) | S4, S5, S6, S8, S11, S12, S15, S17, S21, S22, S23, S25, S26, S27, S30, S31, S32, S33, S34, S35, S36, S37, S38, S39, S40, S41, S42, S45, S46, S47, S48, S52, S53, S57, S58, S60, S61, S62, S63, S68, S70, S72, S73, S74, S75, S76, S77, S79, S80, S81, S83, S84, S86, S87, S92, S93, S98, S99 |
| | False Alarm Rate defined by S85 | S85 |
| | Receiver Operating Characteristic curve (ROC)/ area under the curve (AUC) | S1, S2, S5, S8, S10, S11, S12, S14, S15, S16, S21, S22, S30, S33, S36, S37, S39, S40, S41, S46, S51, S52, S54, S58, S60, S61, S62, S63, S64, S65, S69, S71, S73, S79, S80, S82, S87, S88, S89, S90, S91, S93, S95, S96, S98 |
| | False Negative Rate (FNR) = Missing Rate | S4, S32, S33, S42, S52, S60, S63, S76, S86, S92 |
| | True Negative Rate (TNR) | S42, S52, S60, S83 |
| | Confusion matrix | S4, S21, S60, S62, S89, S90, S92 |
| | Classification Accuracy or Classification rate (CR) | S3, S4, S8, S10, S11, S18, S21, S23, S31, S45, S47, S49, S52, S53, S62, S63, S67, S70, S71, S78, S80, S82, S91, S92, S98 |
| | Precision | S3, S4, S5, S6, S7, S18, S19, S21, S50, S52, S53, S60, S62, S63, S67, S80, S82, S89, S90, S91, S92 |
| | F1, F2(S42) | S3, S4, S5, S6, S7, S18, S19, S21, S24, S42 (F2), S52, S60, S62, S65, S67, S80, S82, S91, S92 |
| | Classification Error | S4, S21 |
| | Matthews Correlation Coefficient (MCC) | S4 |
| Computation Performance | Time (training time/ testing time/ execution time) | S2, S3, S4, S5, S9, S12, S18, S22, S25, S28, S44, S49, S50, S51, S58, S63, S79, S80, S83, S85, S95, S96, S98 |
| | Resource utilization (Storage and computational overhead) | S22, S28, S50, S51 |