Illicit Events Evaluation on the Dark Web Using NSGA-2 Algorithms Based on Energy Consumption

[1*]Romil Rawat, [2]Dr. Anand Rajavat
[1*]Research Scholar, Department of Computer Science Engineering, Shri Vaishnav Institute of Information Technology, Shri Vaishnav Vidyapeeth Vishwavidyalaya Indore, India
[2]Professor and Head, Department of Computer Science Engineering, Director, Shri Vaishnav Institute of Information Technology, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
E-mail: rawat.romil@gmail.com, directorsviit@svvv.edu.in
[*]Corresponding author

Keywords: nsga-2, darkweb, tthe, machine learning, cyber attack, energy consumption, running average power limit

Received: May 21, 2024

The proposed work objective is to adopt the non-dominated sorting genetic algorithm II (NSGA-II), a type of MOEA (multi-objective evolutionary algorithms), to reduce the dimensionality and identify the most relevant features. The objectives and hypotheses tested using NSGA-II and MOEA are to develop predictive models that best fit the problem of traffic analysis, analyse the set of features associated with the dataset, and apply NSGA techniques to reduce the data dimensionality. The work used the datasets and considered the Transnational Terrorist Hostage Event (TTHE) and Counter Trafficking Data Collaborative (CTDC) datasets, which relate to global human trafficking (HTr). The method used is focused on analysing the energy consumption (Econ) with the carbon footprint (CFP) of the proposed algorithm for different values of the main parameters on different hardware platforms, where the quality of each individual is computed based on ML algorithms. The result represents the "F-measure" metric obtained from support vector machines (SVM) and the number of features in order to determine the dimensionality, the execution time, the Econ, and the CFP during the feature engineering process. To estimate performance, the parameter values are computed using SVM.The proposed work concludes that cybercrime employs and represents susceptible behaviour on the OSN Channels network. We identified the threat patterns by using NSGA-II and Econ analysis to establish connections between the variables in the dataset.

Povzetek: NSGA-II algoritem je bil uporabljen in prilagojen za optimizacijo analize nezakonitih dogodkov na temnem spletu.

1 Introduction

In the last few decades, the massive use of the internet has skyrocketed to 5.16 billion users at the beginning of 2023. There is a wide range of applications that provide services through the internet, such as VoIP, gaming, e-mail, financial services, social networking, etc. It is expected that in 2025, more than 5.9 billion users will access these services through mobile devices. The Internet is a means of seducing communication, not only for licit activities but also for illicit ones. The DW, also known as the Dark Net, is the most convenient platform for criminals to carry out illicit tasks, including terrorism [1, 2]. It provides anonymity by hiding IP (Internet Protocol) addresses through a layered encryption model that makes tracing the communication really hard, which stimulates criminal activities [3]. Information mined from internet communications, including DW, about criminal activity is an invaluable source of data. The application of artificial intelligence (AI) techniques to these data helps law enforcement agencies understand, prevent, and anticipate illicit activities. Hostage scenarios are characterised as circumstances in which a terrorist holds one or more people against their will as hostages [4]. These attacks can be motivated by a variety of factors, ranging from expressive goals such as expressing a point of view or religious belief to propaganda-escalation [5] goals such as monetary gain through ransom demands. During hostage situations, there are a variety of options for communication and the usage of an online social platform (OSP), with victims, hostage takers, and media channels all having the ability to comment on and trace the circumstances prior to, during, and after the occurrence of the terrorist event and do identification checks on captives using the OSP and online searches, as shown in the Mumbai [3] incident. They may even pick up their hostages through OSP by tracking movements or personal belongings. Security [45][46][47] agencies use OSP to disseminate official

details to gather background data on hostage-takers relating to political, devout, and personal beliefs, which is publicly available online, in order to facilitate negotiations by better understanding terrorist motivations. Terrorism is a premeditated action carried out by terrorists using secure communication channels such as the DW platform [1]. The hidden world of online traffic, which may be tracked via traffic analysis and pattern recognition, is made up of suspicious activities such as mailing, video streaming, P2P (peer-to-peer) communication and chatting [6], data surfing, chatting, and VOIP (voice over Internet Protocol) [1]. There are multiple techniques used to analyse and categorise DW traffic [7]. However, different challenges arise when working with datasets containing information from communications on different Internet platforms, or DW. Researchers find large and high-dimensional volumes of data, where they usually detect the presence of a high rate of missing values, redundant dates, and unbalanced classes, which affects the precision and reliability of the intelligent model results. Working with high-dimensionality data increases resource needs, and this causes energy consumption and environmental impact to be higher.

According to the International Energy Agency (IEA) [8], energy demand will increase by 40% until 2040. Moreover, energy-related CO2 emissions had the largest rise in 2021, at 36.6 g/t. Predictions for 2050 would reach 32 Gt, far away from the goal of zero emissions objectives, which would cause a 2.5◦C rise in global average temperatures by 2100. Information technology is crucial to achieving sustainability and reducing the carbon footprint. However, they are responsible for 7% of energy consumption (projected to be 13% by 2030) and between 3 and 5% of the carbon footprint [9]. In data centres, 50% of the cost is energy consumption. In this light, the last decade has seen a huge explosion in artificial intelligence (AI) applications. The application of AI is key to fighting against cybercrime [50]. However, it has an important impact on the environment, contributing to CO2 emissions, which can prevent the achievement of sustainability goals. Therefore, it is not enough to use AI to reach sustainability, but it is necessary to address the development of sustainable intelligent models.

Several Methods for Emission Evaluations [32] are available for measuring the energy footprints, like PyRAPL, PyJoules, Perf Command, PowerAPI, AppPowerMeter, PowerMonitor, CodeCarbon, Android Runner, MLCO2, Website Carbon Calculator, and Syspower.

In contrast to other approaches, which call for many independent runs, NSGA-II is predicated on the idea that numerous trade-off solutions can be identified in a single run. One of the main benefits of NSGA-II is that, because its goal is to optimise a scalar objective, a scalar objective local search may be used naturally in each sub-problem.

A set of optimisation methods called MOEAs [43] is used to resolve issues involving several competing objectives. Instead of finding a single ideal answer, these algorithms are made to produce a span of outcomes that show a trade-off between [44] several objectives. MOEA/D (Multi-Objective-Evolutionary-Algorithm-Based-on-Decomposition), SPEA2 (Strength-Pareto-Evolutionary-Algorithm-2) [45], NSGA-II, and many more are examples of common MOEAs. MOEAs have been used in many domains where decision-makers have to balance competing goals, such as engineering design, finance, scheduling, and more.

In contrast to other approaches, which call for many independent runs, MOEA is predicated on the idea that numerous trade-off solutions can be identified in a single run. One of the main benefits of MOEA is that, because its goal is to optimise a scalar objective, a scalar objective local search may be used naturally in each sub-problem.

This work presents an evolutionary approach, particularly a multi-objective approach, using the NSGA-II [10] to carry out the feature engineering process. The objective is twofold: firstly, we want to identify the most relevant features by eliminating those variables that can introduce noise and may influence the assessment; secondly, we study the influence of two important parameters for EAs, particularly the number of generations and the population size, when this process is tackled in two different computing environments. This second aim tries to identify the parameter values that get the best set of relevant features with the lowest energy consumption and environmental impact.

The NSGA [31] optimisation method, which is based on MoEA, is used to discover solutions for two or more competing objectives with specified restrictions.

1.1 NSGA-2 algorithm

- The maximisation of relevance, minimization of redundancy, minimization of feature number, maximisation of accuracy, maximisation of recall, and maximisation of precision are the six objective functions.
- To create offspring, crossover and mutation are applied to the population after initialization. After using this non-dominated sorting, parents and children are joined and categorised by fronts.
- The replacement of a few chromosomes is accomplished by using a randomly generated binary string that has the same length as the decision variables in the issue to be solved.

- The new population is created based on the ordering of the fronts.
- The density of solutions around each solution is used to compute the crowding distance, which is then allocated to each front.
- A tournament shortlisting process is used to choose the children of the following iteration. Finally, crossover and mutation processes result in the birth of a new creation.

To use a genetic algorithm to address multi-objective issues, the objective functions must be transformed into a single objective function. This is changed by assigning weight to each objective function. NSGA-II is a multiobjective problem-solving evolutionary algorithm. Deb [1], [12], and his colleagues proposed this method [15]. The key difference between the GA and the NSGA-II is how individuals are picked for the next generation, while the rest of the stages are nearly identical. The population may be sorted in GA, and individuals may then be chosen for the next generation. Because there are several features for sorting in multi-objective issues, the population cannot be sorted. Many techniques provide weight for each fitness function, reducing multi-objective issues to single-objective problems.

However, it causes objectives to be overlooked. To tackle this difficulty, the NSGA-2 algorithm employs non-dominated sorting. Individuals are assigned to distinct Pareto's based on their rank in non-dominated sorting. The objective is to reduce or increase the fitness of the individuals on the pareto front. Individuals who have not been overpowered by others make up the Pareto front. The concept is that all people are initially compared to one another. If no one else can control an individual, that person is put in front of the others. Then, except for the people who are in front, repeat the procedure for the rest of the people. Individuals who are not controlled by others are put in the second pareto at this phase. This process will be repeated until all of the people are in a pareto. Individuals will be compared using the dominance notion. The notion of dominance is demonstrated using an example below.

If there are two objective functions, F1 and F2, and four individuals, P, Q, X, and Y, and the following conditions:

Cost (F1 (Q)) < Cost (F1 (P)) < Cost (F1 (Y)) < Cost (F1(X))
Cost (F2 (X)) < Cost (F2 (P)) < Cost (F2 (Y)) < Cost (F2(Q))

P dominates Y since it outperforms Y in both goal functions. Other people, on the other hand, do not dominate one another. Because P, Q, and X are not dominated by another person, they are assigned to the front pareto, whereas Y is assigned to the second pareto. For the following generation, the front pareto persons will be chosen. If there is enough room in the following generation for everyone in the next pareto, they will be relocated next to the front pareto, which was moved earlier. The crowding distance value is utilised when only a limited number of individuals may be picked in a pareto for the following generation due to space constraints. Because everyone has the same rank, a crowding distance value is used to choose the best replies from a pareto. The quality of people in the same pareto is defined by the crowding distance. When one person occupies more space than another, the latter has a lower crowding distance.

1.2 Importance for identifying the most relevant features

Reduces training time and Less data means quicker algorithms. A variable's relevance for the present model and prediction is shown by how much each feature adds to the model's prediction using the feature function. Reducing excessive fitting and reducing duplicate data decreases the likelihood that judgements will be based on noise. Improves Accuracy Better modelling accuracy results from fewer misleading data points. Shorter training times and faster algorithms come from less data.

1.3 Current energy and environmental landscape in the world.

Energy consumption (Econ) produces CO2, the main greenhouse gas released. Greenhouse gases raise the Earth's temperature and upset sensitive ecosystems because they are present in the atmosphere. Over the past ten years, data centres have improved their electrical efficiency; experts estimate that electricity only contributes to about 15% of a data centre's emissions, including smart home systems that contribute significantly to CO2 emissions.

1.4 Impact of artificial Intelligence (AI) techniques on the economy and environmental issues.

AI may benefit the ecosystem as well as harm it. While businesses employ AI to boost manufacturing productivity and save energy costs, training AI requires a lot of energy. AI-powered Econ solutions seek to improve production efficiency by predicting energy use and, in particular, machine usage and failure.

1.5 Energy consumption (Econ) and the carbon footprint's importance for AI techniques

AI's Econ is a trade-off between performance and efficiency that is necessary to calculate its CFP and evaluate it against other technologies. AI-based solutions can help with the transition to new energy sources. By using energy supply forecasts, we can find regions where there is potential to increase the usage of solar energy, thereby lowering emissions.

1.6 Background and need of current research

- For determining the emissions on different machines, use the Econ (energy consumption) measure CFP (running average power limit).
- To implement the NSGA-2 algorithm and calculate the Pareto features.
- To calculate the efficiency based on support vector machines (SVM) for the system.

1.7 Research challenges and motivations behind selecting the bi-objective NSGA-II algorithm

- Datasets contain multiple features that might not be important and contain missing values. It becomes difficult to evaluate optimal features for best result generation.
- Unused and indirect dataset features put a burden on processing and increase the energy consumption rate.
- The proposed work has done the analysis to find the missing values in datasets and generated the optimal and confined values for processing.

1.8 Novelty and key contribution

- The proposed work focused on the economy (energy consumption) of a different system for the players involved in cybercriminal activities.
- These studies are beneficial in increasing knowledge of how the cybercrime emission rate is analysed in Google Colab Cloud and in laptop environments.
- The purpose is to understand and quantify the impact that data analysis of cybercrime activities has on energy consumption and carbon footprints.
- A sustainable approach to analysis is presented to discover consistent markers, patterns, and explanations of phenomena that lead to a deeper understanding of cyber threat practices.
- The proposed work created the code and checked the run-time environment and carbon emissions.
- The work is based on light-weight coding code on the Colab platform and the Lab Top platform, which is easy to configure and modifiable and generates different carbon emission values.
- Less research is done in this domain.

1.9 Need for research

- Fitness looks for the trade-off between the" F measure" metric obtained from Support Vector Machines (SVM) and the number of features in order to reduce the dimensionality, the execution time, the energy consumption, and the carbon footprint during the feature engineering process.

1.10 Findings of the study

- Emissions on Different Machines for Calculating Consumption Values: Running Average Power Limit.
- Pareto front evaluation for datasets.
- To evaluate the performance, the parameter values are computed using CNN, LSTM, SVM, and Random Forest algorithms.
- Build a multi-objective scheduling problem optimisation model to execute scientific workflows.
- multi-objective optimisation based on the Pareto optimal method. NSGA-II is used to solve the model.
- We built a cloud with 40 VMs and used a scientific workflow with 100 tasks, then created two matrices (100x40) for the threat and attribute.

1.11 Organization of paper

The remainder of the article is structured as follows: The literature review is presented in Section II, the proposed work is highlighted in Section III, the implementations are shown in Section IV, Section V represents results and outcomes, Section VI shows the discussion, Section VII represents the conclusion, and finally, the future work is shown in Section VIII.

2 Literature review

The feature engineering process has been addressed using different methods, including EA approaches, to eliminate those features that are statistically irrelevant for classification and prediction tasks and to improve the effectiveness of the intelligent models. In [6], the authors underlined significant security [48][49] issues and cutting-edge solutions for IoT networks. Since the filtering approach produces average scores on dataset classes and correctly categorises class labels, it might result in a misleading FS, as seen in [11]. Consequently, it is possible that a feature that is essential for class identification is missed or overlooked. [12] presented an Intrusion Detection System (IDS), which includes a genetic algorithm to identify the most crucial features based on the information theory provided. [13] tackled the feature selection process for a networking dataset by emphasising the optimisation of hyper-parameters for ensemble methods. The authors integrate a swarm intelligence optimisation algorithm with ensemble methods.

In [14], the authors proposed a multi-objective approach using NSGA-II to generate rules for IDS and prevent the effect of a feature from being ignored in subsequent generations. [15] introduced a feature selection scheme called Multi-Objective Evolutionary Feature Selection (MOEFS) that leverages an approach to determine the optimal feature subset for the CICIDS2017 dataset.

In this [16], we used DDOS (denial-of-service attack) based on tack detection techniques for threat classification in DW plat form for tracing the ill-defined events. The work [17] focused on the analysis of darknet markets involved in the trading of contract criminal hirings and the trading of drugs using machine learning (ML) classifiers. The [18] approach is tested on the KDD'99 dataset and the Telecom Bretagne France DDoS dataset on real-time networks.

[19] employed four goals in their work. They used three benchmark datasets to evaluate their suggested methods. They utilised a RF classifier along with an attribute selection approach to evaluate the attainment of the proposed technique. In [1], an NSGA-II strategy for FS specific to IDS is given. This approach focuses on the class imbalance issues of classifiers, which enhances NSGA-II by utilising reference points in order to decrease computing complexity and increase classifier accuracy. On 3-benchmark datasets—NSL-KKDD, KDD-99, and Cure-KDD—the performance of their strategy was assessed using the Jaccard Index. [19] created a hybrid approach based on information theory and evolutionary algorithms. They proposed a multi-objective evolutionary convolutional neural network for IDS in fog environments, where the detection performance and the model complexity of CNN are defined as two objectives to be optimised by a modified MOEA/D. [18] proposed an integrated scheme to filter traffic data to lessen its complexity. This scheme includes an NSGA-II approach to identify the best subset of features, though with execution time constraints.

NSGA-II is a fast and efficient multi-objective evolutionary algorithm (MOEA) [10], which is applied to search for optimal solutions to real problems. The solutions to these problems of ten depend on more than one, and sometimes they have contradictory objectives. NSGA-II has been widely applied to a wide range of different problems [5, 20, 21, 14, 1].

It was not so many years ago that energy consumption when dealing with optimisation problems had become a special subject of interest. Regarding energy awareness when bio-inspired algorithms are applied, different studies have been recently published [22, 23]. These studies highlighted the idea that energy consumption should be optimised and considered part of the solution quality. [24] focused on Natural Language Processing (NLP), where authors encouraged researchers to prioritise both computationally efficient hardware and algorithms. They urged researchers to conduct a cost-benefit analysis by comparing different models to prioritise the most energy-efficient ones.

In [25], the authors presented a review of different methods to estimate energy consumption when machine learning is applied. [26] developed a framework to estimate the real-time energy consumption and carbon emissions of machine learning. They presented results for reinforcement learning algorithms. The authors in [27] analysed performance metrics such as energy consumed by cloud systems to assess the effective resource utilisation and running environment of diverse resources using genetic algorithms. [28] provided a survey of energy-aware scheduling techniques used in modern high-performance computing. [29] evaluated the effect of supervised ML on the energy and memory consumption of intrusion detection systems.

To our knowledge, we have not found studies that analyse energy consumption based on the parameterization of a MOEA, particularly NSGA-II, when dealing with a real problem to improve energy efficiency. The below Table 1 shows about the Summary Table in Related Works.

Table 1: Summary table in related works

References	Method	Parameters
[43]	• Based on an intrusion detection system (IDS) that uses a hybridisation of the NSGA-2 approach and deep learning technology. • Identify distributed denial of service (DDoS) threats in Internet of Things (IoT) networks.	• Using CISIDS2017 datasets on DDoS threats, the experimentation was carried out with a high-performance computer (HPC). • Obtained an accuracy of 99.03% with a 5-fold diminution in training time.

[44]	• A credit card fraud diagnostic dataset was produced by examining the time intervals of individual credit card usage to develop rule-based classifiers. • Easily interpretable using the ENORA and NSGA-II approaches.	• The model's categorisation efficiency with NSGA-II is 94.244% and ENORA = 93.236%, correct categorisation ratios (ACC).
Proposed Work	• The proposed work concludes that cybercrime employs and represents susceptible behaviour on the OSN Channels network. • We identified the threat patterns by using NSGA-II and Econ analysis to establish connections between the variables in the dataset.	• We created several Pareto-front subsets of chosen features, each including one or more chosen characteristics. • With the given approach, it minimised the total count of features from 50 to 11, with a maximum accuracy of 99.51 percent

3 Proposed work

This paper presents a feature engineering approach for assessing cyber-security-related events. As a defensive measure, traffic analytics may be used to track down patterns of interaction and learn vital facts about potential cybersecurity risks. To minimise the dimensionality and pinpoint the most important characteristics, this work used NSGA-II, one of the most frequently used MOEAs. Moreover, the study includes the evaluation of two of the main parameters for a MOEA: the number of generations and the population size for tracking energy consumption and carbon emissions. As we previously mentioned, the main goal is to identify the most relevant features in the target datasets and those parameters that obtain the best trade-off between solution quality and energy efficiency. However, energy consumption also depends on the hardware platform where the experiments are run. In this light, we test the experiments using two different platforms, Google Colab and a laptop, which will be described in the 3.3 section.

To carry out this process, the general scheme presented in Figure 1 is followed. This scheme presents several phases, which are described below.

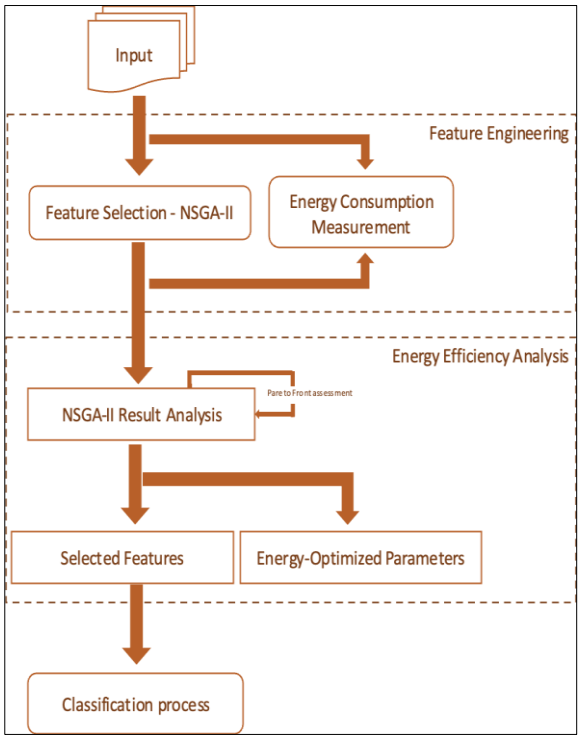

Figure 1: General scheme for the feature engineering process and the identification of the best energy-efficient parameters to obtain the best relevant features.

a. Input phase

This phase is responsible for providing the input data to the next phase, the FS one. The proposed approach uses the datasets from the Transnational Terrorist Hostage Event (TTHE) [7] and Counter Trafficking Data Collaborative (CTDC) [30], which are related to international human trafficking threats (HTT). Data in both datasets that we are collab orating with are taken from DW, and using the TOR browser to track host-generating DW crime and malevolent behaviour, as well as covertly utilising criminal markets, such as drug counterfeiting, money laundering, organ trafficking, cyber terrorism, extremism and staffing, criminal hiring, the gambling and armaments industry, extortion, and HTr. Correlations and useful data are also looked at for future design development.

i. Transnational terrorist hostage event (TTHE):

The TTHE dataset represents cyber terrorist attacks and events of mass destruction, as well as threats, which spans the years 1978 to 2018 [3]. Based on information from media sources, OSP, and anonymous DW platforms, TTHE incorporates up to 50 factors for each of its 1,974 events.

- The four sorts of hostage scenarios that are documented for a global sample are kidnappings, sky jacking and non-aerial hijacking.
- Some of the most important factors include the count of hostages, demands made by the kidnappers, terrorist attributes during mediation, the quantity of ransom requested, other sorts of demands, the length of the crisis, and the result of bargain.
- The beginning and end of the hostage scenario, the size of the attack force, the count of victims, casualties, and weapons used, and the logistical outcome are all included in our hostage data collection. TTHE is related to International Terrorism, the Attributes of Terrorist Events (ITERATE) [31].
- TTHE is used to find the logistical and negotiation success elements for hostage deployments using fascinating details of hostage occurrences throughout history. The target property is normalized in the 0 to 1 range [12, 7], whereas the characteristics are normalized in the -1 to 1 range.
- TTHE shows the characteristics of hostage situations over time and may be used to determine what factors influence the success of hostage missions in terms of logistics and negotiation.

b. Counter trafficking data collaborative (CTDC):

The CTDC dataset contains information on human trafficking records (HTr) and cases across the world, along with exploitation details.

- The first worldwide data centre on human trafficking is CTDC [30] which collects information from organisations all across the world. Data on HTr included on TTHE-Dataset have historically been challenging for analysts, researchers, practitioners, and policymakers to easily acquire.
- The absence of trustworthy, high-quality data is widely acknowledged as one of the major obstacles to designing focused counter-trafficking actions and assessing their effectiveness.
- Data on human trafficking is sometimes an extremely contentious topic.
- This dataset includes details on 48.8K HTr victims, such as the cause, the method of restraint, the origin and destination, as well as other factors.
- The CTDC provides un-identified, individual-level data on human trafficking victims, including case records and trends in human trafficking. In order to prevent human trafficking and represents numerous ground-breaking elements.
- It provides unprecedented detail on human trafficking incidents and vital insights into victim sociodemographic profiles and human trafficking dynamics to develop targeted interventions and improve survivor assistance.
- It publishes harmonised and synthesised individual case data from multiple counter-trafficking organisations worldwide.

i. Choice of TTHE and CTDC datasets in the context of energy consumption and carbon footprint analysis

The TTHE represents cyber terrorist attacks and events of mass destruction, as well as threats, and the CTDC dataset contains information on human trafficking records (HTr), and the CTDC represents individual-level data on human trafficking victims, including case records and trends in human trafficking, which helps to prevent human trafficking and represents numerous ground-breaking elements.

The purpose of selecting these datasets is to work on cyber threats generated by terrorists relating to trafficking cases. The dataset contains multiple features relating to threats, threat actors, and events, but still, the proposed research is focused on susceptible behaviour on the OSN Channels network, so NSGA-II helps to generate useful features.

The purpose of the work is more focused on understanding the behaviour of terrorist events at the OSN network by analysing energy consumption and carbon footprint analysis to establish and understand connections between the variables in the dataset. This will help to model future datasets for the training of ill-defined events using energy evaluation parameters and online behaviour.

The dataset contains enormous quantities of data for many different properties; however, none of this data is helpful because certain aspects are meaningless and don't provide any important information. Modelling this data may result in extra costs for processing and storage, as well as longer execution times. By applying dimension reduction to the dataset, these overheads may be reduced to a minimum, enabling the analysis of energy consumption (Econ) and carbon footprint (CFP) with target attributes for focused research.

c. Feature engineering phase

This process is tackled using EAs, particularly an MOEA approach with NSGA-II [10], one of the de facto standard multi-objective optimisation algorithms. Mono-objective genetic algorithms consider a single objective, and many problems can be addressed using such an approach, but most real problems need multiple objectives to be evaluated. In this work, the aim is to reduce the data dimensionality, which will improve the speed of AI models during the training and productivity phases. Next, we identify the most energy-efficient NSGA-II parameters to reduce energy consumption and carbon footprint emissions when feature engineering is addressed.

d. Multi-objective optimisation function

This proposal defines a decision variable $x \in X$ in a multi-objective feature extraction framework. Variable x defines a set of features that belong to the dataset addressed. The evaluation process consists of computing the multi-objective function F as the number of features and the F-measure metric resulting from SVM as a supervised ML algorithm. Thus, the best solution will correspond with low values of the number of features and high values of the F-measure metric, which are conflicting objectives.

The multi-objective algorithm evolves to minimise the number of features and maximise the F-measure metric obtained after the individuals' evaluation with the SVM supervised learning algorithm. As a result, a set of not-dominated solutions is obtained after the number of generations is parameterized, which reflects the best set of attributes to identify cybercrime events. This set of solutions is the is the Pareto front, an approximation of the Pareto optimal front. The concept of non-dominance means that solutions belonging to the same front may perform better on one objective but worse on another.

One of our objectives is to pinpoint the most sustainable parameter values, so we selected a range of values for the two main parameters of NSGA-II: the number of generations and the population size. Therefore, our feature engineering process is launched multiple times, each time focusing on different parameter values. For each test, the energy consumption and the carbon footprint impact were computed to analyse which parameter values reached the best trade-off between the solutions' quality and the energy consumption.

e. System architecture

The data has been preprocessed and normalised to the {O 1} range. To minimise the dimensionality of the data, the normalised data is put into the NSGA-II [13]. The reduced data from the previous stage is sent into the ML models for classification as input data. The proposed system architecture is depicted in Figure 2.

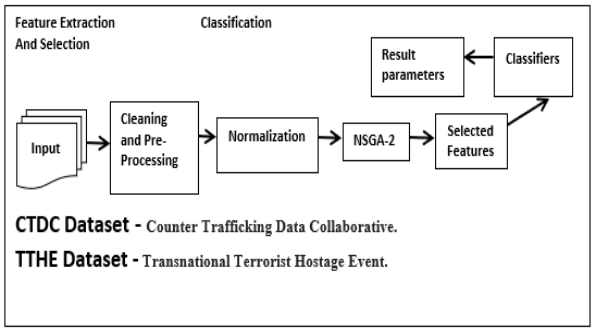

Figure 2: Proposed system architecture

f. Taxonomy of TOR Traffic data

In Figure 2, the data created by TOR (the Onion Router) [14] and [15] is categorized. The data is sent from the TOR browser to the DW router using a key and hash value generated by the TOR directory with relay nodes, which keeps the data's identity hidden. The traffic data is analysed, and useful information and patterns are analysed for further design development in order to trace host-generating DW crime [16] and malicious behaviour, as well as secretly operating illicit [17] markets of drug trafficking, money laundering, organ trafficking [16], cyber terrorism, radicalization and recruitment [18], contract crime induction, illicit gambling and weapon business, cyber extortion, human trafficking and abuse,

and the selling of stolen antiques. The classification of data generated by TOR (the Onion Router) [12] is shown in Figure 3.

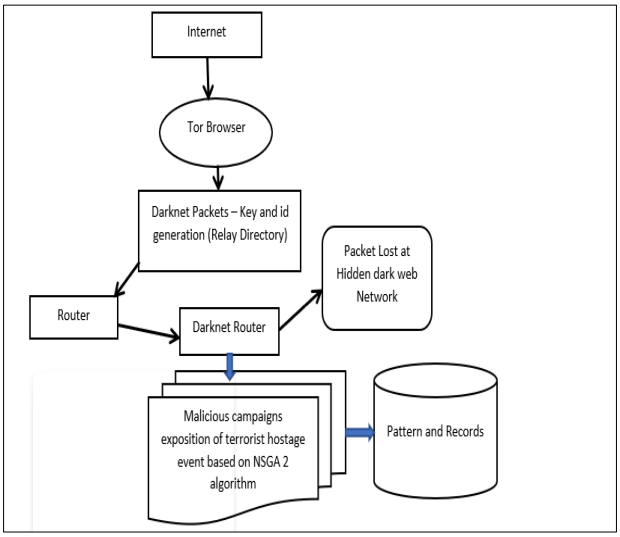

Figure 3: Classification of data generated from TOR (The Onion Router) [12]

4 Implementations

The set of experiments designed has three different goals: (i) tackling the detection of dangerous actions on the DW; (ii) minimising the dimensionality and identifying the best set of features to accurately reach the first goal; and (iii) determining the most sustainable algorithm parameter values to achieve the two previous goals. Code is run in two different computing environments: a laptop and the Google Colab-Cloud platform.

Individuals can utilise their current computer hardware to tackle much larger issues at a lower extra cost by using a technique to discover the best virtual machine (VM) placement in a cloud environment based on parallel computing. Large computational tasks may thus be tackled more affordably by combining the processing capacity and memory of several computers, which also enhances resource consumption and lowers server energy use. As a result, distributed systems need efficient parallel architectures and approaches that can handle large amounts of data. Method parallelization is therefore becoming increasingly essential since it makes it possible to use such techniques using the computational capacity found in huge data centres, particularly in cloud computing configurations.

The virtual machine environment (computer nodes) is where a possible solution can be evaluated. It has a parallel environment with" X" virtual machines, where" X" individuals are run in parallel to speed up the algorithm. For evaluating individuals with the SVM technique using this computing environment, we need to store individuals in a matrix with a specific size, and for that, this computing environment needs individuals to be stored in a matrix, and then every individual will be assigned to a VM, where the SVM technique is run.

The Matrix throughput and resource management solution aims to provide predictable VM performance with the highest level of effectiveness. In order to accomplish this, Matrix uses clustering techniques with probability estimates to forecast how new tasks will function in a virtualized environment. It then selects an appropriate VM type and dynamically modifies the VM's resource configuration as needed. The Matrix assigns VM resources to applications in a way that minimises workload despite attaining high performance, precisely predicting how an emerging workload will operate on various cloud VM instances. The Matrix additionally oversees a variety of cloud environments, applications, and workloads while preserving intended performance and maximising cloud cost-effectiveness.

As previously mentioned, a MoEA strategy with NSGA-II is applied that takes energy considerations into account. As a defensive measure, traffic analytics may be used to track down patterns of interaction and learn vital facts about potential cybersecurity risks. This approach considers the datasets from the CTDC and TTHE, which are related to international HTT.

This study has been carried out using the methodology presented in this section and following the data flow shown in Figure 1. The data that has been previously pre-processed and normalised to the O-1 range is provided as input to the NSGA-II to carry out the proposed feature extraction and selection phase. The same set of experiments has been launched for both datasets using Table 2. The below table represents the parameters for specific values and impact on results.

- with the lowest parameters for population size, number of generations, and mutation probability.
- The crossover rate begins to decline. The number of generations and the size of the population both affect how long it takes for the algorithm to execute.
- to determining which generation achieves the highest level of fitness.

Table 2: NSGA-II parameters' values

NSGA-II parameter	Parameter values
Number of generations	100, 200 and 300
Population size	128, 256, 512 and 1024

The algorithm is run in the cloud using Google-Colab and on a laptop using PyCharm to estimate the power consumption and the carbon footprints, while the algorithm is running to identify the best set of attributes to identify cyber security events.

Using the bi-objective NSGA-II strategy implemented, the algorithm is launched according to the settings shown in 1 to identify the most sustainable parameter values for the number of generations and population sizes. For each experiment, the individual's quality is determined based on

(i) the outcomes of the SVM algorithm and (ii) the number of features. The NSGA II algorithm is depicted schematically in Figure 4. The different stages are described as follows:

1) Initialize Population: P_i
2) Random Population generation: Count- C;
3) Objective Values Evaluation;
4) As Per Pareto -Sort: Assignment of Rank (Level);
5) Child Population Generation;
6) Selection of Binary Tournament ;
7) Process of Recombination and Mutation;
8) *For i=1 to m , do*
   a. *For every Parent and Child in the Population do*
   b. Based on Pareto – Sort : Assignment of Rank (level);
   c. Generation of nondominated solution sets;
   d. Determining Crowding distance;
   e. Starting by the First front until M individuals: next generation solution using Loop(inside) addition;
9) *End*
10) By high crowding distance : Lower front points selection;
11) Next generation Creation :
12) Selection of Binary Tournament;
13) Process of Recombination and Mutation;
14) *End*

Figure 4: NSGA-II- Pseudocode for Features representation

While the algorithm is running, the power consumption and carbon footprint measurements are quantified and collected. Using the EmissionTracker library, tracker start() is used to define the initial point of the code to be analysed, and tracker.stop() stops the estimation process. As a result, these measurements are stored for later analysis to identify the most sustainable parameter values compared to the quality of the solutions obtained on both addressed platforms. The results attained are described in the following section. Attributes selected by the algorithm, those included in the pareto front- *Location_start and Type_of_incident.*And for CTDC the pareto front's *isSexualExploit and recruiterRelationFamily* attributes were chosen by the algorithm.

a. Algorithm

The experiments have been implemented at Google Colab and on personal laptops. Multiple iterations have been done, and values are observed as given below.

i. Login at Google Colab
ii. Import Dataset: TTHE and CTDC
iii. Run the NSGA-2 algorithm.
iv. Calculate efficiency and energy parameters using Codecarbon and EmissionsTracker.
v. Run ML algorithms and calculate system efficiency.

And for CTDC, the pareto front's isSexualExploit and recruiterRelationFamily attributes were chosen by the algorithm.

i. Parameter evaluation

The measurements about the Econ were taken on Google colab to study the performance of the Econ and the CFP.By using the code carbon package, the emission file evaluated multiple parameters, as given here.

'Duration', 'emissions','emissions_rate', 'cpu_power','gpu_power','ram_power', 'cpu_energy','gpu_energy', 'ram_energy','energy_consumed','cpu_count','cpu_model', 'gpu_count', 'gpu_model', 'ram_total_size'.

5 Result and outcomes

The evaluation of the proposed method is conducted in a Python environment by Colaboratory-Google Research. Multi-objective optimisation produces results as a set of Pareto- functions according to the objective functions defined as maximising or minimising. Code is implemented at the Google Colab Platform for the cyber crime emission rate and is analysed in the cloud environment. We built a cloud with 40 VMs and used a scientific workflow with 100 tasks, then created two matrices (100x40). Each attribute is counted as a virtual machine, and each complete row is defined as an individual task chromosome.

Attributes selected by the algorithm for TTHE are those included in the pareto front: location_start and type_of_incident.

- Build a multi-objective scheduling problem optimisation model in which the threat and attribute to execute scientific workflows are considered.
- multi-objective optimisation based on the Pareto optimal method. NSGA-II is used to solve the model.
- We built a cloud with 40 VMs and used a scientific workflow with 100 tasks, then created two matrices (100x40) for the threat and attribute.

DWMA network data is obtained and normalized. This data is put into the NSGA-II algorithm, which has a number of key requirements for FS methods. The classifier is trained on both attacked and unaffected data, and the evaluation is then validated. In the range O to 1, the DWTA network data is acquired, preprocessed, and normalized. This normalised data is sent through the NSGA-2 [13] to reduce the data's dimensionality. We've thought about the most significant objectives for implementing the algorithm. As input data, the model receives the reduced data created in the previous stage.

a. Target hardware platforms

This study addressed two different hardware platforms to identify the most sustainable one. The cloud platform,

Google Colab, and a laptop have been chosen. Our aim is also the identification of the most efficient from an energy-efficient perspective and environmental impact in terms of carbon emissions.

Google Colab is a freeware cloud-based platform that includes a Jupyter notebook interface and free access to CPUs, graphics processing units (GPUs), and tensor processing units (TPUs) [32]. In Colab, there is a 2.2GHz processor, 13GB of memory, a 33GB hard drive, and an Nvidia Tesla K80 GPU. On the one hand, Google Colab offers fast accessibility to note books with free GPUs and TPUs and requires no setup or prior software deployment. On the other hand, it's a rather restricted workplace because only the pre-installed Python module may be used by machine learning experts. The programmer cannot add their own Python package and begin executing the code. As a result, the framework can offer basic tools but is unsuitable for specialised use.

Laptop platform: This hardware platform includes a processor with an 11th Gen Intel (R) CoreTM i5-1135G7 and 2.40GHz–2.42GHz. The system contains installed RAM of 8.00 GB (7.75 GB usable). The system type is a 64-bit Windows 7 operating system with an x64-based processor.

b. Measure energy consumption methods

The open-source Python library CodeCarbon [33] is used for estimating the CFP, the quantity of carbon dioxide equivalents (CO2e), and the energy consumption in Google Colab and laptop environments. These estimations are carried out by measuring the power consumption of the GPUs, CPUs, and RAM components. CodeCarbon can be used on NVIDIA GPUs with NVIDIA Management Library (NVML) support and on CPUs supporting Intel's Running Average Power Limit (RAPL) [34] interface or Inter Power Gadget [35]. The EmissionTracker library provides the functionality to track emissions according to power consumption and location-dependent carbon intensity. The massive applications of AI to develop AI models need huge hardware and software resources for the training phase and also for the production phase. As a result, AI models have a considerable impact on carbon emissions and energy consumption. For now, the design of AI models needs to be evaluated to identify the more efficient and sustainable parameters both during the training and production phases.

The proposed work offers an efficient method for identifying the best set of features in cybersecurity-related frameworks using a multi-objective approach. During the execution of the algorithm, power consumption and carbon emissions are tracked to analyse the best algorithm parameters to improve sustainability targets. The next section details the experiments carried out.This section details the results obtained for the different experiments carried out using the TTHE and CTDC datasets, according to the specifications described in Section 4. We structure the section depending on the dataset we deal with. For each dataset, we present results reached using the tackled platforms. In all cases, we analysed the solutions provided by the algorithm implemented, the energy consumption and carbon footprint for each parameter value, and the quality concerning the dimensionality reduction as shown in Table 3.

Table 3: F Measure values Count- TTHE and CTDC dataset

Features	F Measure (TTHE)	F Measure (CTDC)
3	87.23	85.21
7	89.58	88.76
8	89.98	90.78
9	90.21	90.87
11	91.68	89.65
14	94.44	93.98
16	94.89	95.02
22	95.24	96.12
26	96.21	97.43
31	96.89	97.05
36	97.11	97.91
41	97.87	98.07

DWMA network data is obtained and normalized. This data is put into the NSGA-II algorithm, which has a number of key requirements for FS methods. The classifier is trained on both attacked and unaffected data, and the evaluation is then validated. In the range O to 1, the DWTA network data is acquired, preprocessed, and normalized. This normalised data is sent through the NSGA-2 [13] to reduce the data's dimensionality. We've thought about the most significant objectives for implementing the algorithm. As input data, the model receives the reduced data created in the previous stage.

The TTHE dataset is used for evaluation. NSGA-2 (used for optimal features) Extraction, based on refined data generated by the classifiers (CNN, LSTM, SVM, and RF), [2] [3][4][21][25] shows the result metrics (precision, recall, accuracy, F-measure)**.** Table 4 below shows a description of classifier algorithms.

c. *Parameters setting for the methods*

The below algorithm represents the parameter setting for the model [46]. Figure 5 shows the algorithm of the proposed work.

- ***Start energy consumption (E) tracking (from codecarbon import EmissionsTracker)***
  #Measuring the energy consumption and carbon footprints of the system
- **Load Datasets (D)**
  #with labelled data
  - D_1-THE
  - D_2- CTDC
- ***Start of the NSGA-2 Algorithm:***
  - ***For (i to n)***
  - Take different values for population size (P*)* and *number of generations (G).*
    - $P_{1...n}$: { i to n: 64, 64 256, 512}
    - $G_{1...n}$: { i to n: 100, 200, 300, 400, 500}
    - For every P_i, select G_i.
- **End of the NSGA-2 Algorithm:**
  - The ***result*** will generate the **nsga2_results.xlsx** file with features set (S):
  #containing the minimum set of features (attribute value and count) and the quality of the solutions: the number of features and the metric.
- **S for New dataset: Apply SVM Algorithm and Measure Value (F1_measure).**
  #F1-measure gives the quality of selected attributes for cyber-attack behavior analysis on different machines.
- **Calculate parameters (accuracy, precision, recall value).**
- **Stop Start energy consumption Tracing, and the output will generate an emission.xls file**.
  - Energy Consumption and carbon footprint values.

Figure 5: Algorithm of proposed work

d. Accuracy diagram (obtain ----- F - measure as result)

The values of several algorithms employed in research are summarised in the table below. Based on experiment data, the NSGA-2 values precision at 99.51%, recall at 99.53%, and F1-Score at 98.48%. The comparison of parameter values of different ML algorithms is shown in Table 4.

Table 4: Comparison of parameter values

Method	Precision	Recall	Fl-Score
CNN	96.69%	98.52%	97.10%
LSTM	98.96%	98.81%	98.93%
SVM	95.50 %,	97.45%	96.98%
Random Forest	94.64 %.	96.76%	95.98%

In this study, the NSGA-2+ classifier algorithms achieve the best results and outperform the other algorithms in terms of accuracy ratio. The suggested method's first step is to reduce the data by utilising
NSGA-2 to do FS. As a consequence of applying this to the data, we were able to come up with a number of solutions.

The comparison of different numbers of characteristics such as precision, recall, and Fl-Score is shown in Table 2. It's worth noting that the set of eleven features acquired has the best accuracy, recall, and FI-Score values of 98.96%, 98.81%, and 98.93%, respectively. Figure 6 depicts the graph of the F1 Score vs. Recall Graph.

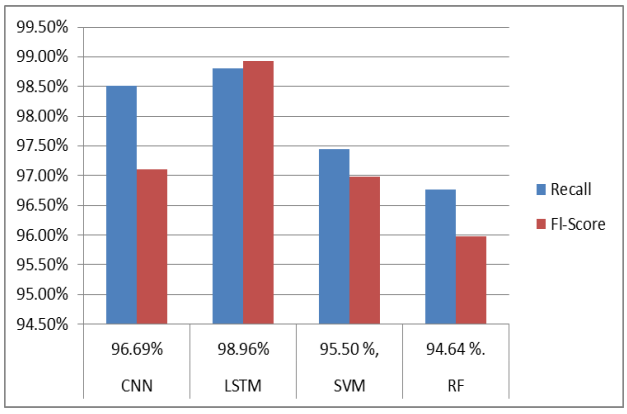

Figure 6: F1 Score v/s recall graph

e. Parameter evaluation

The measurements about the Econ were taken on Google Collaborate to study the performance of the Econ and the CFP. Figures 6 and 7 below [36] [1] show the methods for emission evaluation. Table 5 shows the Pareto front evaluation resulting from the application of NSGA-II to the CTDC dataset. And Table 6 shows the Pareto front evaluation resulting from the application of NSGA-II with the TTHE dataset and the Table 7 represents about the Comparison of Results with Existing Techniques. By using code carbon package, the emission file evaluated multiple parameters as given here: duration, *emissions, emissions_rate, cpu_power, gpu_power, ram_power, CPU_energy, gpu_energy, ram_energy, Energy_consumed, cpu_count, cpu_model, gpu_count, gpu_model, and ram_total_size*.

The findings demonstrate that a bi-objective algorithm fitness function aids in the development and training of highly effective host-state classifiers through the use of a minimised subset that outperforms baseline models in terms of prediction performance.

- After execution of the code. The file (nsga2_results.xlsx) will be generated, containing generated attribute values for SVM to calculate the F-measure.
- The emission file (emission.csv) will also be generated after execution, containing Energy consumption parameters as given in Figure- 8 and Figure-9.
- The code-carbon package is used to calculate the carbon emissions. The process is iterated to analyse the emission attributes as shown in Figure-7

```
[codecarbon INFO @ 10:43:01] Energy consumed for RAM : 0.000059 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:43:01] Energy consumed for all CPUs : 0.000381 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:43:01] 0.000441 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:43:16] Energy consumed for RAM : 0.000079 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:43:16] Energy consumed for all CPUs : 0.000509 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:43:16] 0.000588 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:43:31] Energy consumed for RAM : 0.000099 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:43:31] Energy consumed for all CPUs : 0.000636 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:43:31] 0.000735 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:43:46] Energy consumed for RAM : 0.000119 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:43:46] Energy consumed for all CPUs : 0.000763 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:43:46] 0.000882 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:44:01] Energy consumed for RAM : 0.000139 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:44:01] Energy consumed for all CPUs : 0.000890 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:44:01] 0.001028 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:44:16] Energy consumed for RAM : 0.000158 kWh. RAM Power : 4.753040313720703 W
[codecarbon INFO @ 10:44:16] Energy consumed for all CPUs : 0.001017 kWh. Total CPU Power : 30.5 W
[codecarbon INFO @ 10:44:16] 0.001175 kWh of electricity used since the beginning.
[codecarbon INFO @ 10:44:31] Energy consumed for RAM : 0.000178 kWh. RAM Power : 4.753040313720703 W
```

Figure 7: Process is iterated to analyze the emission attributes

S.No.	Parameters/Machine	Machine-1 (Colab) (in kWh)	Machine -2 (Laptop) (in kWh)
1	CPU Energy Consumption	0.000601	0.000634
		0.000355	0.000413
		0.000532	0.000622
2	Ram Energy Consumption	0.000067	0.000073
		0.000020	0.000031
		0.000040	0.000047
3	Total Energy Consumption	0.000668	0.000712
		0.000040	0.000051
		0.000592	0.000623
4	Duration (milliseconds)	121.409384	36.91793108
		121.418682	36.96499989
		122.316543	37.47792816

Figure 8: Emissions on different machines (For TTHE dataset)- econ measure (CFP)- running average power limit

S.No.	Parameters/Machine	Machine-1 (Colab) (in kWh)	Machine -2 (Laptop) (in kWh)
1	CPU Energy Consumption	0.004781	0.005143
		0.004604	0.004945
		0.000354	0.004236
2	Ram Energy Consumption	0.000535	0.001923
		0.000515	0.006545
		0.000040	0.001312
3	Total Energy Consumption	0.005316	0.006345
		0.005119	0.006269
		0.003151	0.003998
4	Duration (milliseconds)	219.6417499	303.7312383
		127.0086961	264.4582637
		201.7587388	238.8465366

Figure 9: Emissions on Different Machines (For CTDC dataset)- Econ Measure (CFP)- Running Average Power Limit

Table 5: Pareto front evaluation resulting from the application of NSGA-II with the CTDC dataset. NG, PS and NF are the number of generations, population size and number of features

Run	NG	Pop. size	NF	F-Measure	Energy Consumption	Carbon Footprint	Reduction Percent
1	100	128	5	90.22	0.000247	0.2	20
2		1024	4	89.96	0.000712	0.08	8
3		256	4	89.35	0.000540	0.06	6
4		512	4	88.87	0.000592	0.06	6
5	200	512	5	94.89	0.000634	0.08	8
6		1024	5	89.93	0.000493	0.06	6
7		128	5	89.92	0.000551	0.06	6
8		256	4	89.91	0.000608	0.06	6
9	300	1024	4	91.35	0.000618	0.06	6
10		512	4	89.82	0.000631	0.08	8
11		256	4	89.24	0.000673	0.08	8
12		128	3	87.92	0.000611	0.06	6

Table 6: Pareto front evaluation resulting from the application of NSGA-II with the TTHE dataset. NG, PS and NF are the number of generations, population size and number of features

Run	NG	Pop. size	NF	F-Measure	Energy Consumption	Carbon Footprint	Reduction Percent
1	100	1024	3	90.12	0.000545	0.06	6
2		512	3	89.98	0.000515	0.06	6
3		256	3	89.58	0.000519	0.06	6
4		128	8	87.23	0.000734	0.18	18
5	200	1024	4	95.24	0.000669	0.09	9
6		512	4	94.89	0.000694	0.09	9
7		256	3	94.44	0.000599	0.06	6
8		128	4	91.68	0.000610	0.09	9
9	300	1024	5	97.87	0.000686	0.11	11
10		512	4	97.11	0.000660	0.09	9
11		256	3	96.89	0.000556	0.06	6
12		128	3	96.21	0.000523	0.06	6

Table 7: Comparison of results with existing techniques

Reference	Technique Name	Methods	Features/Limitations
[37]	related to intangible assets.	worked on the standard protocol for quantifying greenhouse gas emissions and computing the carbon footprint.	Limited to two years' work and design to implement this methodology. limited to static attributes only.
[38]	to incorporate scientific findings regarding climate change into their applications.	Platform-agnostic Open Carbon Footprint Framework provides interfaces for developers	Limited to fuel-efficient mobile GPS applications, the iPhone only suggests the greenest fuel-efficient route to the user.
[39]	to assist automation for products for SMEs in Taiwan.	Carbon footprint calculation	the carbon footprint of the final product is automatically analysed by the tool.
[40]	to define the main sources of climate impacts in the IT industry with an adaptation for software companies.	model of the Greenhouse Gas Protocol	Only the framework is presented for the protocol to be applied to the software industry.
[41]	The algorithms are at the meta-learning phase, which involves assessing loss after each cycle and accurately modifying the weights as a result of the DL stage's rising complexity.	comprehensive overview of carbon emissions based on deep learning	More ecologically friendly are simpler machine-learning algorithms, as they may frequently provide comparable outcomes with less energy and processing capacity.
[42]	The calculator tool that author designed and implemented calculates the carbon footprint.	web page tool based on server location	This toll will assume all page views are done on a smartphone only.
Proposed Work	Illicit Event Evaluation on the Dark Web Platform Based on Dataset Values	Used NSGA-2 and machine learning algorithms for calculating and analysing energy consumption.	The proposed work created the code and checked the run-time environment and carbon emissions. The work is based on light-weight coding code on the Colab platform and the Lab Top platform, which is easy to configure and modifiable and generates different carbon emission values. Less research is done in this domain.

f. *Comparison of NSGA-II with other evolutionary algorithms*

- When comparing NSGA-II to Strength Pareto Evolutionary Algorithm 2 (SPEA-II), it was found that NSGA-II was the best tested algorithm.
- It also showed the best CPU run time across all test scales.
- By accounting for elitism, NSGA-II promotes excellent convergence and quickens it.
- Better elitism and variety are preserved with NSGA-II.
- Through domain analysis, NSGA-II generates Pareto-optimal solutions rather than a single solution. These solutions are arranged in an effective manner.

6 Discussion

The proposed work is focused on DW traffic analysis using the TTHE and CTDC datasets by calculating the energy consumption with carbon footprint (CFP) on different hardware platforms, and the quality of each individual is computed based on ML algorithms. The present work

represents the lightweight process and architecture for understanding the environment and behaviour of terrorist and criminal data using NSGA and machine learning algorithms with respect to energy consumption parameters. Figure 1 shows the flow of work for the feature engineering process and the identification of the best energy-efficient parameters to obtain the best relevant features. The Fig. 2 shows the proposed architecture. Fig. 3 displays the classification of data generated from TOR (the Onion Router) with the proposed algorithm. Table 1 shows the NSGA-II parameters' values taken for feature evaluation. Figure 4 displays the NSGA-II pseudocode for feature representation. Table 2 shows the F Measure Values Count (TTHE and CTDC datasets. Table 3 displays the comparison of parameter values; Fig. 5 shows the F1 Score vs. Recall Graph. The Figure 6 shows about the process is iterated to analyse the emission attributes, Figure 7 shows the emissions on different machines (for the TTHE dataset). Econ Measure (CFP): Running Average Power Limit, Figure 8 displays the emissions from different machines (for the CTDC dataset): Econ, Table 4 shows the Pareto front evaluation resulting from the application of NSGA-II to the CTDC dataset. NG, PS, and NF are the number of generations, population size, and number of features, and the Table 5 shows the Pareto front evaluation resulting from the application of NSGA-II to the TTHE dataset. NG, PS, and NF are the number of generations, population size, and number of features and the Table 6 shows about the Comparison of Results with Existing Techniques.

Threat pattern catalogues are security pattern catalogues used to assess security use cases and provide a threat taxonomy. They contain vague event goals and the word and phrase patterns that terrorist groups employ for communication at OSN. The OSN postings and messages are evaluated and mapped using dataset features in order to find threat trends. The technology collects logs related to words and phrases that match and are based on dataset features in order to assess thorough threat intelligence and information sources.

7 Limitations

- The system performance is deeply dependent on the dataset quality and NSGA algorithm, and based on this, the result will represent its behaviour.
- Econ also depends on hardware quality and configuration

8 Conclusions

FS in the TTHE dataset and CTDC is optimised using a MOEA approach using the NSGA-2 and ML algorithms. The traffic assessment method is used to track down communication patterns and learn vital details about potential security risks. To reduce data complexity and improve the speed of DW traffic analysis, feature engineering is focused. The NSGA-II and MOEA techniques were used in this study to reduce dimensionality and pinpoint the most important traits. This approach searches a wider region for solutions, which enables the algorithm to provide a high count of Pareto-efficient answers. We created several Pareto-front subsets of chosen features, each including one or more chosen characteristics. With the given approach, it minimised the total count of features from 50 to 11, with a maximum accuracy of 99.51 percent, which was the best result. The datasets are assessed simultaneously using the NSGA-2 and SVM algorithms. The Emissions on Different Machines Econ Measure (CFP) (RAPL) is calculated to analyse the emission rate.

9 Future work

- The experimental analysis section would benefit from incorporating statistical analyses, such as the Friedman test and Nemenyi test, to highlight differences among the methods used.
- In the future, the emission rate will be calculated on multiple machines using a set of datasets to better analyse the attributes and behaviours of cybercrime on a cloud platform.

Abbreviation table

Abbreviation	Details
MoEA	multi-objective evolutionary algorithm
FS	feature selection
DW	dark web
Econ	Energy consumption
NSGA-II	non-dominated sorting genetic algorithm-ii
TTHE	transnational terrorist hostage event
CTDC	counter trafficking data collaborative
HTr	human trafficking
CFP	carbon footprint
AI	artificial intelligence
IP	internet protocol
OSP	online social platform
VOIP	voice over internet protocol
SVM	support vector machines
P2P	peer-to-peer
IEA	international energy agency
Econ	energy consumption
IDS	intrusion detection system
MOEFS	multi-objective evolutionary feature selection
CICIDS2017	intrusion detection evaluation dataset
DDOS	denial-of-service attack
ML	machine learning

KDD	knowledge discovery in databases
CNN	convolutional neural network
NLP	natural language processing
HTT	human trafficking threats
ITERATE	international terrorism, the attributes of terrorist events
EA	evolutionary algorithms
GPUs	graphics processing unit
TPUs	tensor processing units
RAM	random-access memory
CO2e	carbon dioxide equivalents
NVML	nvidia management library
RAPL	running average power limit
VM	virtual machine
NG	number of generations
PS	population size
NF	number of features
CNN	Convolution Neural Network
LSTM	Long Short-Term Memory Network

References

[1] Aksu, D., & Aydin, M. A. (2022). MGA-IDS: Optimal feature subset selection for anomaly detection framework on in-vehicle networks-CAN bus based on genetic algorithm and intrusion detection approach. *Computers & Security*, *118*, 102717. https://doi.org/10.1016/j.cose.2022.102717

[2] Enoch, S. Y., Mendonça, J., Hong, J. B., Ge, M., & Kim, D. S. (2022). An integrated security hardening optimization for dynamic networks using security and availability modeling with multi-objective algorithm. *Computer Networks*, *208*, 108864. https://doi.org/10.1016/j.comnet.2022.108864

[3] Kim, W., George, J., & Sandler, T. (2021). Introducing transnational terrorist hostage event (TTHE) data set, 1978 to 2018. *Journal of Conflict Resolution*, *65*(2-3), 619-641.https://doi.org/10.1177/0022002720957714

[4] Guermouche, A., & Orgerie, A. C. (2022). Thermal design power and vectorized instructions behavior. *Concurrency and Computation: Practice and Experience*, *34*(2), e6261.https://doi.org/10.1002/cpe.6261

[5] Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V. E., ... & Zhukov, L. E. E. (2022, December). Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. In *Doklady Mathematics* (Vol. 106, No. Suppl 1, pp. S118-S128). Moscow: Pleiades Publishing.https://doi.org/10.1134/S1064562422060230

[6] Bayerl, P. S., Akhgar, B., Brewster, B., Domdouzis, K., & Gibson, H. (2014). Social media and its role for leas: review and applications. *Cyber Crime and Cyber Terrorism Investigator's Handbook*, 197-220.https://doi.org/10.1016/B978-0-12-800743-3.00016-5

[7] Das, S. P. (2022). Hostage-Taking, Ransom, and Negotiations. In *Economics of Terrorism and Counter-Terrorism Measures: History, Theory, and Evidence* (pp. 481-503). Cham: Springer International Publishing.https://doi.org/10.1007/978-3-030-96577-8_12

[8] International Energy Agency. (2022). *World energy outlook 2022*. International Energy Agency. https://www.iea.org/reports/world-energy-outlook-2022https://doi.org/10.1016/S0360-5442(99)00045-6

[9] Karamchandani, A., Mozo, A., Gómez-Canaval, S., & Pastor, A. (2024). A methodological framework for optimizing the energy consumption of deep neural networks: a case study of a cyber threat detector. *Neural Computing and Applications*, 1-42.https://doi.org/10.1007/s00521-024-09588-z

[10] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, *6*(2), 182-197.DOI: 10.1109/4235.996017

[11] Oh, O., Agrawal, M., & Rao, H. R. (2011). Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers*, *13*, 33-43.https://doi.org/10.1007/s10796-010-9275-8

[12] Xia, T., Qu, G., Hariri, S., & Yousif, M. (2005, April). An efficient network intrusion detection method based on information theory and genetic algorithm. In *PCCC 2005. 24th IEEE International Performance, Computing, and Communications Conference, 2005.* (pp. 11-17). IEEE.DOI: 10.1109/PCCC.2005.1460505

[13] Chohra, A., Shirani, P., Karbab, E. B., & Debbabi, M. (2022). Chameleon: Optimized feature selection using particle swarm optimization and ensemble methods for network anomaly detection. *Computers & Security*, *117*, 102684.https://doi.org/10.1016/j.cose.2022.102684

[14] Tamimi, A., Naidu, D. S., & Kavianpour, S. (2015, October). An Intrusion detection system based on NSGA-II Algorithm. In *2015 Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic (CyberSec)* (pp. 58-61). IEEE.DOI: 10.1109/CyberSec.2015.20

[15] Panigrahi, R., Borah, S., Pramanik, M., Bhoi, A. K., Barsocchi, P., Nayak, S. R., & Alnumay, W. (2022). Intrusion detection in cyber–physical environment using hybrid Naïve Bayes—Decision table and

multi-objective evolutionary feature selection. *Computer Communications*, *188*, 133-144.https://doi.org/10.1016/j.comcom.2022.03.009

[16] Bu, S. J., Kang, H. B., & Cho, S. B. (2022). Ensemble of deep convolutional learning classifier system based on genetic algorithm for database intrusion detection. *Electronics*, *11*(5), 745. https://doi.org/10.3390/electronics11050745

[17] Rawat, R., Mahor, V., Chirgaiya, S., Shaw, R. N., & Ghosh, A. (2021). Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021* (pp. 671-681). Springer Singapore.https://doi.org/10.1007/978-981-16-0749-3_53

[18] Dimolianis, M., Pavlidis, A., & Maglaris, V. (2021). Signature-based traffic classification and mitigation for DDoS attacks using programmable network data planes. *IEEE Access*, *9*, 113061-113076.DOI: 10.1109/ACCESS.2021.3104115

[19] Chen, Y., Lin, Q., Wei, W., Ji, J., Wong, K. C., & Coello, C. A. C. (2022). Intrusion detection using multi-objective evolutionary convolutional neural network for Internet of Things in Fog computing. *Knowledge-Based Systems*, *244*, 108505.https://doi.org/10.1016/j.knosys.2022.108505

[20] Jaisingh, W., Nanjundan, P., & George, J. P. (2024). Machine Learning in Cyber Threats Intelligent System. In *Artificial Intelligence for Cyber Defense and Smart Policing* (pp. 1-20). Chapman and Hall/CRC.https://doi.org/10.1016/j.knosys.2022.108505

[21] Zidi, I., Issaoui, I., El Khediri, S., & Khan, R. U. (2024). An approach based on NSGA-III algorithm for solving the multi-objective federated learning optimization problem. *International Journal of Information Technology*, 1-13.https://doi.org/10.1007/s41870-024-01801-5

[22] Yadav, S., Hashmi, H., & Vekariya, D. (2024). Mitigation of attacks via improved network security in IOT network environment using RNN. *Measurement: Sensors*, 101046.https://doi.org/10.1016/j.measen.2024.101046

[23] Rekeraho, A., Cotfas, D. T., Cotfas, P. A., Bălan, T. C., Tuyishime, E., & Acheampong, R. (2024). Cybersecurity challenges in IoT-based smart renewable energy. *International Journal of Information Security*, *23*(1), 101-117.https://doi.org/10.1007/s10207-023-00732-9

[24] Alwaisi, Z., Soderi, S., & De Nicola, R. (2023, October). Detection of Energy Consumption Cyber Attacks on Smart Devices. In *International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures* (pp. 160-176). Cham: Springer Nature Switzerland.https://doi.org/10.1007/978-3-031-50051-0_12

[25] Brehler, M., Camphausen, L., Heidebroek, B., Krön, D., Gründer, H., & Camphausen, S. (2023). Making Machine Learning More Energy Efficient by Bringing It Closer to the Sensor. *IEEE Micro*.DOI: 10.1109/MM.2023.3316348

[26] Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., ... & Zhao, L. (2024). Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*. https://doi.org/10.48550/arXiv.2401.00625

[27] Jaiprakash, S. P., Arya, H. K., Gupta, I., & Badal, T. (2023). Energy Optimized Workflow Scheduling in IaaS Cloud: A Flower Pollination based Approach. *Authorea Preprints*.DOI: 10.22541/au.168594506.69867134/v1

[28] Kocot, B., Czarnul, P., & Proficz, J. (2023). Energy-aware scheduling for high-performance computing systems: A survey. *Energies*, *16*(2), 890. https://doi.org/10.3390/en16020890

[29] Baidoo, C. Y. M., Yaokumah, W., & Owusu, E. (2023). Estimating Overhead Performance of Supervised Machine Learning Algorithms for Intrusion Detection. *International Journal of Information Technologies and Systems Approach (IJITSA)*, *16*(1), 1-19.DOI: 10.4018/IJITSA.316889

[30] Macy, R. J., Klein, L. B., Shuck, C. A., Rizo, C. F., Van Deinse, T. B., Wretman, C. J., & Luo, J. (2023). A scoping review of human trafficking screening and response. *Trauma, Violence, & Abuse*, *24*(3), 1202-1219.https://doi.org/10.1177/15248380211057273

[31] Hu, Y., & Man, Y. (2023). Energy consumption and carbon emissions forecasting for industrial processes: Status, challenges and perspectives. *Renewable and Sustainable Energy Reviews*, *182*, 113405.https://doi.org/10.1016/j.rser.2023.113405

[32] Shanbhag, S., & Chimalakonda, S. (2023, May). An exploratory study on energy consumption of dataframe processing libraries. In *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)* (pp. 284-295). IEEE.DOI: 10.1109/MSR59073.2023.00048

[33] Raffin, G., & Trystram, D. (2024). Dissecting the software-based measurement of CPU energy consumption: a comparative analysis. *arXiv preprint arXiv:2401.15985*. https://doi.org/10.48550/arXiv.2401.15985

[34] Zhang, X., Ding, A. A., & Fei, Y. (2023, October). Deep-Learning Model Extraction Through Software-Based Power Side-Channel. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)* (pp. 1-9). IEEE.DOI: 10.1109/ICCAD57390.2023.10323806

[35] Olu-Ajayi, R., Alaka, H., Owolabi, H., Akanbi, L., & Ganiyu, S. (2023). Data-driven tools for building energy consumption prediction: A review. *Energies*, *16*(6), 2574. https://doi.org/10.3390/en16062574

[36] Sallou, J., Cruz, L., & Durieux, T. (2023). EnergiBridge: Empowering Software Sustainability through Cross-Platform Energy Measurement. *arXiv preprint arXiv:2312.13897*. https://doi.org/10.48550/arXiv.2312.13897

[37] Loyarte-López, E., Barral, M., & Morla, J. C. (2020). Methodology for carbon footprint calculation towards sustainable innovation in intangible assets. *Sustainability*, *12*(4), 1629.https://doi.org/10.3390/su12041629

[38] Rahman, F., O'Brien, C., Ahamed, S. I., Zhang, H., & Liu, L. (2011). Design and implementation of an open framework for ubiquitous carbon footprint calculator applications. *Sustainable Computing: Informatics and Systems*, *1*(4), 257-274.https://doi.org/10.1016/j.suscom.2011.06.001

[39] Chen, J. L., Chen, W. C., & Kuo, A. (2016, September). Developing carbon footprint calculation software for display industry in Taiwan. In *2016 Electronics Goes Green 2016+(EGG)* (pp. 1-7). IEEE.DOI: 10.1109/EGG.2016.7829863

[40] Sipilä, A., Partanen, L., & Porras, J. (2023, November). Carbon Footprint Calculations for a Software Company–Adapting GHG Protocol Scopes 1, 2 and 3 to the Software Industry. In *International Conference on Software Business* (pp. 442-455). Cham: Springer Nature Switzerland.https://doi.org/10.1007/978-3-031-53227-6_31

[41] Rahimbakhsh, H., Kohansal, M. E., Tarkashvand, A., Faizi, M., & Rahbar, M. (2022). Multi-objective optimization of natural surveillance and privacy in early design stages utilizing NSGA-II. *Automation in Construction*, *143*, 104547.https://doi.org/10.1016/j.autcon.2022.104547

[42] Peng, Y., Wang, Y., Chen, H., Wang, L., Luo, B., Tong, H., ... & Chen, S. (2024). Carbon reduction potential of a rain garden: A cradle-to-grave life cycle carbon footprint assessment. *Journal of Cleaner Production*, *434*, 139806.https://doi.org/10.1016/j.jclepro.2023.139806

[43] Roopak, M., Tian, G. Y., & Chambers, J. (2020, January). An intrusion detection system against DDoS attacks in IoT networks. In *2020 10th annual computing and communication workshop and conference (CCWC)* (pp. 0562-0567). IEEE.DOI: 10.1109/CCWC47524.2020.9031206

[44] Dirik, M. (2022). Predicting credit card fraud using multipurpose classification based on evolutionary rules. *Security and Privacy*, *5*(5), e239.https://doi.org/10.1002/spy2.239

[45] Sabir, A., Ali, H. A., & Aljabery, M. A. (2024). ChatGPT Tweets Sentiment Analysis Using Machine Learning and Data Classification. *Informatica*, *48*(7).https://doi.org/10.31449/inf.v48i7.5535

[46] Liu, Y., & Pan, B. (2024). Profit Estimation Model and Financial Risk Prediction Combining Multi-scale Convolutional Feature Extractor and BGRU Model. *Informatica*, *48*(11).https://doi.org/10.31449/inf.v48i11.5941

[47] Rawat, R., & Rajavat, A. (2024). Perceptual Operating Systems for the Trade Associations of Cyber Criminals to Scrutinize Hazardous Content. *International Journal of Cyber Warfare and Terrorism (IJCWT)*, *14*(1), 1-19. DOI: 10.4018/IJCWT.343314

[48] Rathi, S. K., Keswani, B., Saxena, R. K., Kapoor, S. K., Gupta, S., & Rawat, R. (Eds.). (2024). *Online Social Networks in Business Frameworks*. John Wiley & Sons. DOI:10.1002/9781394231126

[49] Vyas, G., Vyas, P., Muzumdar, P., Chennamaneni, A., Rajavat, A., & Rawat, R. (2024). Extracting and Analyzing Factors to Identify the Malicious Conversational AI Bots on Twitter. *Conversational Artificial Intelligence*, 71-83. https://doi.org/10.1002/9781394200801.ch5

[50] Taiwo, G. A., Saraee, M., & Fatai, J. (2024). Crime Prediction Using Twitter Sentiments and Crime Data. *Informatica*, *48*(6). https://doi.org/10.31449/inf.v48i6.4749