International Journal of Safety and Security Engineering
Vol. 15, No. 11, November, 2025, pp. 2261-2273
Journal homepage: http://iieta.org/journals/ijsse

Detection and Classification of Darknet Traffic Using an Enhanced LocalKNN Algorithm

Ekram H. Hasan[1*], Hussein K. Almulla[2], Omar A. Dawood[3], Mohammed Khalaf[4]

[1] Department of First Grades Teacher, College of Basic Education – Haditha, University of Anbar, Ramadi 31001, Iraq
[2] Electronic Computer Center, University of Anbar, Ramadi 31001, Iraq
[3] Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Ramadi 31001, Iraq
[4] Department of Computer Science, University of Al Maarif, Ramadi 31001, Iraq

Corresponding Author Email: ekram.habeeb@uoanbar.edu.iq

https://doi.org/10.18280/ijsse.151107

Received: 23 July 2025
Revised: 24 September 2025
Accepted: 15 November 2025
Available online: 30 November 2025

Keywords:
classification, darknet traffic, dark web, data preprocessing, LocalKNN algorithm, machine learning, traffic detection, deep web

ABSTRACT

Using the dark web is inevitable due to the wide range of anonymous services that are provided, which can be accessed via the Onion Router (TOR) and Virtual Private Network (VPN). Anonymity creates an environment where malicious activities occur out of reach of law enforcement. Such activities have huge implications for society's safety and organizations' cybersecurity environment. Therefore, this paper tackled these challenges by proposing a darknet traffic detection and categorization model that can flag users who can access the darknet. The proposed model utilizes different techniques in machine learning, starting from pre-processing (data cleaning, encoding, and balancing) that prepares data before selecting the top 50 best features, using Chi-Square, which can then be used to train a classifier model built based on local K-Nearest Neighbors (KNN) and City and Hamming as distance metrics. The improved K-Nearest Neighbors with Local Metric Induction (LocalKNN) algorithm calculates a local metric as an additional step for each classified object. During the process of classifying test objects, classifiers use global metrics to find a large set of nearest neighbors, which is then used to derive a new local metric set. Finally, a locally induced metric is used to select the nearest neighbors. The model shows a great potential result in detecting traffic with an accuracy of 99.34% and a categorization accuracy of 92.34%.

1. INTRODUCTION

The internet includes three layers: the Surface Web, the Deep Web, and the dark web. The dark web, a subset of the Deep Web, is frequently used for saving and retrieving confidential information online, which is considered a hidden, untraceable Internet layer. Its content is purposefully concealed and inaccessible to regular web browsers [1].

However, several events have shown the covert use of this platform for criminal and unlawful activity. So, this layer is associated with numerous forms of criminal conduct, such as gambling and purchasing. The Internet is not equal among the three layers; around 96% of the Internet is made up of the deep web and the dark web. The dark web and its unlawful activity have gained notoriety in recent years due to a variety of social factors and anonymous accessing using decentralized nodes of specific network organizations (such as Invisible Internet Project (I2P) or Onion Router (TOR)) [1, 2]. The users' anonymity makes them useful in both legal and unlawful circumstances. This provides protection for those who want to use the dark web for illegal activities with little possibility of being discovered. These websites serve as a starting point for web users who require privacy. Additionally, it provides encryption for users to avoid tracking their online footprint by the service provider or government. Due to its anonymous nature and the complexity of the dark web, as well as the use of advanced techniques to hide activities and bypass detection. Therefore, there is a need to provide a method for detecting dark web traffic and extracting relevant data and useful information to help law agencies in investigating these activities that take place on the dark web [3]. Reaching the darknet can be the first step that cybercriminals take; therefore, detecting and preventing such traffic can alleviate the risk of using the darknet illegally. For this reason, creating detection mechanisms is a must to protect networks and regular users. Many researchers have conducted investigations that have led to the proposed different approaches to detect such traffic. Using machine learning was mostly the best option due to its ability to extract features, find correlations among these features, and detect patterns. Using a learning-based approach led many afford into creating reliable datasets that can be used in evaluating the proposed learning-based approach. With such an opportunity, researchers can focus their attention on constructing learning models that can achieve higher accuracy and reduce the time needed to draw a conclusion. Using a time-consuming model can reduce its usability because the server or firewall processes massive traffic in a short time. So, choosing a fast and reliable model can improve the likelihood

of adopting such a model. Using the nearest neighbor-related model is popular due to its effectiveness and relative speed compared to other models.

This paper aims to provide a study on the impact of the dark web and investigate an approach that utilizes machine learning to detect dark web traffic, and investigates several types of nearest neighbor variant algorithms on the challenge of detecting and categorizing darknet traffic. The paper proposes a detection and categorization model depending on the use of K-Nearest Neighbors with Local Metric Induction (LocalKNN), which is a variant version of K-Nearest Neighbors (KNN) with the use of local metric indication. The structure of this paper is as follows. Section 2 surveys some previous related studies. Section 3 explains the dark web and the most important components, characteristics, applications, activities, risks, threats, and implications of the dark web on cybersecurity. Section 4 includes the proposed approach for detecting and categorizing darknet traffic. Section 5 explains the experiment and discusses the results of performance measures. Finally, this paper is concluded in Section 6.

2. LITERATURE REVIEW

The elevated risk of the dark web made it a topic of interest and gained significant attention from researchers. This section will review some of the previous related research conducted on studies and applications for monitoring and detecting the dark web.

In 2019, Schäfer et al. [4] presented BlackWidow, which is a modular system that is used to highly automate the monitoring services of dark web services and amalgamate the collected data into a single framework for further analysis. Technically, the micro services (Docker) were the base on which BlackWidow relies. BlackWidow was able to collect data and relevant information in the domain of fraud monitoring and cybersecurity for an extended period. The collected data and its relationship, which were extracted from posts, are represented by a large knowledge graph used by security analysts for further investigation and exploration.

In 2019, Han et al. [5] focused on detecting spatiotemporal patterns that occur due to malware infection activity. Finding such a pattern can help in detecting malware during its process of spreading the infection over the network. Spreading malware can involve using multiple hosts and ports, which mostly have similar patterns (they are called synchronized). The proposed approach focused on detecting that pattern and finding such synchronization, which can be a step forward in detecting potential malware activities. This approach adopted the machine learning method "Graphical Lasso" to estimate the synchronization of spatiotemporal patterns from darknet traffic data. Also, to evaluate and examine the early detection of malware activity.

In 2021, Habibi Lashkari et al. [6] proposed an approach depending on deep learning for darknet traffic detection is converting the traffic into a 2D image and uses a deep image network for classification. The approach was evaluated on different datasets (ranging from the year 1989 to 2017) and reported the highest accuracy of 86%. Using CNN-LSTM as a classifier and XGB as a feature selection was applied on datasets and reported a detection accuracy of 96% and darknet traffic categorization accuracy of 89%.

In 2022, Mohanty et al. [7] proposed an approach that involves using multiple algorithms to improve performance by utilizing a stacking ensemble design to combine three base learning algorithms (Random Forest, KNN, and Decision Tree) to detect darknet traffic. To increase robustness, the proposed approach used a two-layered autoencoder-based system, which consists of a detector and a denoiser, as a defense mechanism against adversarial attacks. It was evaluated on the CIC-Darknet- 2020 dataset and reported an accuracy of 98.89%.

In 2023, Almomani [8] used a stacking ensemble with different algorithms. The approach uses three base learners (Random Forest, Neural Network, and SVM) whose predictions pass the second level to make decisions using logistic regression. It was evaluated using the CIC-Darknet2020 dataset and reported an accuracy of 97% in detecting darknet traffic.

Table 1 summarizes the benefits and limitations of each application in the above-mentioned studies. It is worth noting that our application outperformed all the mentioned techniques through the high accuracy of identification and classification.

Table 1. Comparison between existing methods

Citation	Year	Technology Used	Advantages	Limitation
[4]	2019	Blackwidow system (dark web monitoring framework)	- Highly automated and scalable - Real-time cyber threat intelligence extraction	- Requires initial manual setup to identify target forums - High resource consumption for crawling (e.g., Puppeteer) - Relies on machine translation (possible semantic inaccuracies) - Short lifetime of many dark web forums limits long-term tracking - Limited accuracy
[5]	2019	The GLASSO engine is based on a sparse structure learning algorithm known as "graphical lasso"	- Real-time malware detection - The model generated alerts for three types of activities. i.e., survey scans sporadically, focused traffic, and cyberattacks - GLASSO engine can precisely and automatically determine random cyberattacks like network scans	- The GLASSO engine has some false negatives

			in real time	
[6]	2020	Deep Image Learning approach: traffic features converted to grayscale images and analyzed by 2D Convolutional Neural Networks (CNN).	- Provided classification accuracy (~86%) for darknet traffic - Using image-based representation reduced the need for large labeled datasets	- Accuracy is not perfect for all applications - Long training time and high computational cost - Difficulties in handling unstructured data or Encrypted
[7]	2022	Robust Stacking Ensemble (RF, KNN, DT + Meta-classifier)	- Provide an accuracy score of 98.89% in the darknet traffic identification. Furthermore, an accuracy of 97.88% is reached in the case of darknet traffic flow characterization. - Solid against adversarial attacks (BIM, FGSM)	- Requires Higher computational cost - Deals with varied and large datasets - Limited generalization, requires wider validation - No deep packet inspection (limits on privacy and performance)
[8]	2025	Modified Stacking Ensemble Learning (base: NN, RF, SVM; meta: Logistic Regression)	- Provide high accuracy in darknet traffic classification. The system achieved greater than 99% accuracy in the training phase, while in the testing phase, it implemented an accuracy of 97% - Deals with unknown and known darknet attacks - Suitable for larger datasets	- Initial assessment is limited to specific datasets - Needs to be examined on a wider range of attack types

3. DARK WEB

The websites that are hidden or not searchable on the Internet are collectively referred to as the "dark web". It resides on a particular network that is hidden from view on a regular web browser. It is often involved in illegitimate acts, including selling illicit products and services. It can also be used for lawful objectives, like exchanging private information and communicating anonymously. It is a private web where users' names and their location are hidden by encryption technology that routes information through numerous servers across the world, making it particularly challenging to track down specific people [9]. Accessing the dark web requires special kinds of browsers such as the TOR, I2P, FreeNet, Whonix, and Riffle. While there are many legitimate uses for the dark web as well, it is true that criminal activity on the dark web is more prevalent than on the Deep Web or the regular Internet. Legitimate applications encompass activities like utilizing TOR to examine reports of domestic violence, political oppression, and other crimes that have grave repercussions for individuals who bring them to light [10].

In 2011, Ross William Ulbricht contributed to finding the most well-known electronic bazaar on the dark web called "Silk Road". Silk Road is an online market that focuses on the trading of drugs, fraud, piracy services, malware, passports, hacked media, and other services and goods. The site was shut down by the FBI in September 2013, and Ulbricht was taken into custody in October of that same year. According to the US Federal Court, he was given a life term in jail in 2015 after earning more than $13 million through trading items and commissions for services, while the website generated over 1.2 billion dollars in revenue from 2013 to 2022 [11].

The evolution of the Web has relied on a variety of protocols and tools; the essential components of the dark web are the following [12]:

1. Web browsers, which are used to access the dark web.
2. Encryption technique to encrypt the data for security, which is a crucial component.
3. Identity protection, such as the TOR browser, protects the identity by employing random routes and multiple layers of complicated encryption.
4. Virtual private networks (VPNs) for data transmission.
5. Routing algorithm. It's essential to maintain anonymity when using the dark web. Using a decent VPN in addition to your browser is necessary to maintain your anonymity.

3.1 Onion Router

Distributed and anonymous nodes on several networks are used to access the dark web, such as TOR, which is considered the most essential tool. TOR is a multilayer encryption method that is low-latency, widely available, and represents dark web content. Onion Router gets its name from the fact that it contains several layers, much like an onion [13]. Multi-layer architecture design allows a device to identify only the current and prior devices that the traffic is passing through, without knowing the original source and final destination devices. On each device, the data packet of the current layer is decrypted by the device, which allows it to know the prior and next devices, so it knows where to send it next; however, it is unaware of the main origin or final destination. As a result, data packets cannot be tracked without the network's origin or target. TOR uses addressing methods made of randomly generated keys to maintain the "hidden sites." TOR URLs are typically long and difficult to remember, and the websites' URLs are modified to thwart Distributed Denial of Service (DDoS) assaults and divulge information. Users can access the dark websites from their computer if any of these browsers are installed. Also, due to the structure of TOR, it can be used for innocent purposes, like simply browsing the Internet for daily purposes without being constrained by prohibited content or worried about government or private surveillance [14]. Relying on the "Onion Routing" concept, TOR encrypts user data before sending it via several relays spread throughout the TOR network. To protect the user and conceal their identity, layered encryption is employed. It functions by bypassing connections made through a series of intermediate relays from the user's computer to the destination. At every relay that comes after, an encoded layer is decoded, and the remaining data is routed to any relay until it reaches its destination server.

The exit relay appears to the target server as the data source. This encoded data is then further encoded, making it only decodable by entering the relay. It indicates that, like an onion, the hierarchical encryption mechanism has covered unique data; in each layer, there is data that it needs to be aware of regarding where it obtained the data and where it is sending it [15].

3.2 Dark web applications and activities

There are several situations in which understanding and using the dark web could be beneficial. This section will survey the most important dark web applications:

- Darknet market commonly referred to as the crypto market, are commercial or tradable sites that are used for selling commodities on the dark web, and they are accessible using specific browsers such as TOR, I2P, or Freenet, which conceal user identity and location. Darknet market serves as a black market because they are used for selling and dealing with malware, illegal substances, forged currency, weapons, illegal pharmaceuticals, stolen bank information, and other illegal items or services, as well as legal goods. According to Gareth Owen from the University of Portsmouth, in December 2014, the darknet market was ranked as the second most visited website on the TOR dark web. Dark wallets are used with Bitcoin-based transactions on the darknet to protect the buyer and seller while maintaining anonymity [16].
- Bitcoin Encouragement: One well-known cryptocurrency is Bitcoin, which is a decentralized electronic payment system. Bitcoin utilizes cryptography for anonymous payments and security. On the P2P (peer-to-peer) Bitcoin network, it may be transmitted directly between users without the need for an intermediary; therefore, a central bank or authority is not needed to facilitate the operation. Users often get Bitcoins by "mining" them, receiving them as payment, or exchanging them for fiat money. Any financial transaction involving Bitcoin is documented in a blockchain, which is a publicly accessible ledger that includes senders' and recipients' addresses. The address only identifies a specific transaction; it does not uniquely identify any specific Bitcoin. Users'addresses are associated with and stored in wallets, which also store the private key. The private key is like a password, a secret number that permits the owner to spend Bitcoins on the associated wallets. Spending these Britons on any transactions requires verifying ownership by verifying a cryptographic signature and the transaction address. Because the wallet and private key are not visible to the public ledger, using Bitcoin offers an extra layer of privacy. Therefore, users of Bitcoin are able to conduct anonymous business transactions, which could encourage some illegal activities [16, 17]. The symmetric key can also be used in blockchain to secure a transaction; however, the algorithm must be lightweight and fast, as the studies [18, 19] have shown.
- Military using anonymity: on the dark web helps prevent opponents from identifying and breaching military systems while they are in the field. The dark web can be used by the military to research the environment in which it operates and to identify behaviors that could endanger soldiers' ability to carry out their duties. The military can utilize TOR software to carry out secret information or secret computer network operations, such as denial-of-service attacks and the takedown of websites, or to intercept and obstruct communications from adversaries. Another use could be psychological operations or military deception, in which the military plants false information on targets and the movements of troops on the dark web in order to be gathered by intelligence or to undermine the narrative of the rebels [17].
- Intelligence Community (IC): The fact that any user trying to download TOR was automatically electronically fingerprinted means that the IC can potentially identify users who think they are untraceable. As a result, the IC uses it as a source of open intelligence [17].

3.3 The dark web underground economy

According to the applications mentioned in the previous section, one of the most well-known features of the dark web is unlawful marketplaces, often known as darknet market, where a variety of illicit goods and services are sold. This creates an "underground economy", which is a hidden network where illegal activities and transactions occur. This economy includes a variety of instruments used in cybercrime. The problem is overly complex due to the fact that some crimes overlap different areas of national security. Public safety, medical care, and finances are all impacted by drug trafficking and dealing. In addition to the tax-related offenses already mentioned, the cryptocurrency marketplaces bolster the shadow and gray economies. Trading counterfeit items causes damage to the companies' reputation and profit by lowering their incomes and damaging the quality of their brands and the consumer confidence. Trading counterfeit electronic payment information, such as credit cards, bank account information, gift cards, and other services, which can include stolen bank accounts, causes severe damages for financial institutions. These crimes not only inflict harm but also undermine public trust in these systems [20].

3.4 Cybercrimes on the dark web

Because the dark web offers a variety of services that facilitate anonymous networking, the dark web is an attractive ground for cybercriminals. The following is a list of the most reported crimes on the dark web.

- Information leakage: A lot of anonymity-supporting networks, like TOR, are helpful tools for law enforcement, activists, and whistleblowers. In same time, hackers can use the dark web as a platform to disclose confidential information [21].
- Proxying: The anonymity provided by the TOR platform makes its users more exposed to attack. The reason is that the hypertext transfer protocol (HTTP), which is normally accessible through regular Uniform Resource Locators (URLs), is missing. As a result, the security measures associated with https are ignored. They must bookmark the TOR page in order to ensure they are on the official website. When a user is tricked into believing they are on the legitimate page using internet proxying, the criminal modifies the link to bring the user to his fraudulent website. Whenever an

individual pays using cryptocurrency, the money ends up going straight to the con artist. On a website like this, a normal transaction involving an untraceable cryptocurrency like Bitcoin will con the user while the hacker moves the funds [21].

- Onion cloning: The working principle of onion cloning is similar to that of the proxy. To steal money from users, the fraudster copies the legitimate website or page and updates the links to send users to the fraud sites [21].
- Bitcoin scam: The most popular cryptocurrency on the dark web is Bitcoin. The Dark Net marketplace often only accepts Bitcoin as a currency that allows the members to stay anonymous while processing transactions. Using cryptocurrency can raise illegal activities and increase Cyber-terrorists in the world of the dark web, even in the case of legitimate use of Bitcoin, money laundering increases [22].
- Hidden Wiki: It is one of the most common locations on the dark web for newbies. There are several connections on it that promote money laundering, murder contracts, cyberattacks, limited chemical bonding, and explosives manufacturing instructions. But the URL to the Hidden Wiki continues changing, just like other dark web sites that require frequent URL changes to avoid detection by law enforcement. Hidden Wiki is the way to browse malevolent intentions and illegal materials in the dark web by offering links to the Deep Web [23].
- Frauds: Financial fraud represents one of the illegal financial transactions within the dark web. InstaCard and Banker and Co., the two most notable dark sites, enable the user to make illegal and untraceable financial transactions. These websites include several URLs that lead to the same page for the user. These illicit transactions are mostly carried out using two methods: 1) hiding the true transaction's origin to launder cryptocurrencies and 2) creating an anonymous payment option, such as a debit card, to be used by a financial institution [21].
- Terrorism: The dark web provides a fertile ground for terrorist groups, as the dark web provides an anonymous network hidden away from the public Internet. For that, it meets the terrorists' needs to conduct financing, recruitment, propaganda, and conspiracy operations and to close their sites easily and without being traced [21, 23].

3.5 Implications of the dark web for cybersecurity practices

Overall, the dark web has had a significant impact on cybersecurity practices by increasing the sophistication of cyber threats, creating underground marketplaces, challenging law enforcement efforts, and facilitating the distribution of hacking tools. This section will highlight the significant impacts of the dark web on cybersecurity [24].

- The military aspect is one of the most important aspects of cybersecurity for any country; therefore, leaking or stealing critical document data in order to sell military hardware blueprints linked to a nation's national security, such as tanks, unmanned aerial vehicles, etc. The material is then given to hackers who attempt to justify the stolen data.
- Terrorists utilize the dark web to finance and obtain weapons.
- Zero-Day Exploits: according to the perspective of national security, the spread of zero-day exploits poses a number of issues, which can lead to the most serious and difficult risks to computers and network systems from malevolent actors who profit from them on the dark web.
- Online anonymity: permits people to freely express themselves without any limitations while keeping their online identity separate from their real identity. With the advent of technologies, such as the dark net, many people with comparable views come together in one network, which allows them to freely express themselves without fear of reprisals or harassment. It is typical for users of these sites to communicate anonymously or behind pseudonyms that they have selected for themselves, which makes it challenging to identify offenders.
- Cybercriminal Activity: Regarding cybersecurity threats, hacking communities are active on dark web platforms where hackers share information and exchange experiences. They also circulate ransomware, malware, hacking tools, and compromised data, and plan massive cyberattacks that resemble organized crime patterns.

3.6 Risk mitigation of the dark web

The structure design of the dark web creates a highly anonymous environment, which creates a challenge for government security agencies and forensic professionals to discover traffic information, such as the origin, physical location, or ownership of devices on the Darknet. Law enforcement agencies utilize a variety of countermeasures and strategies to mitigate the risks posed by the dark web, such as:

- Application of monitoring techniques: Because of the anonymous nature of the TOR network, it is incredibly difficult to monitor the dark web. The Onion Router (TOR) is an untraceable and hard-to-shut-down architecture that creates a playground for criminals to navigate the dark web. This is among the factors driving the intense demand on law enforcement and security organizations to keep an eye on and track activity on the Dark web [25]. Dark web monitoring may help detect crimes and criminal activities by tracking and analyzing communication channels, chatrooms, and emails, which can be a forum to discuss illicit activities to discover illegal extraction [25].
- Dark web criminal and crimes detection: There are methods that might be implemented to improve the safety and dependability of the dark web, such as identifying criminals on the dark web and, if feasible, pinpointing their locations, such as tracking criminal activity through social media, and implementing stronger civil and criminal laws to prosecute offenders. Also, using a special search engine called "MEMEX" that can explore the dark websites that are not indexed by standard search engines. MEMEX applies advanced algorithms to reveal patterns and links in the content on the dark web [26].
- Developing systems for criminal investigations that rely on digital evidence to enable digital criminal law enforcement agencies to get beyond barriers to

apprehend offenders and reduce their violent behavior. The most widely used digital forensic tools are Autopsy and Encase, which can analyze the file system's disk, disk figures, partitions, and password retrieval [26, 27].

- Enhanced Law Enforcement Efforts: Governments have been focusing their efforts on improving law enforcement agencies' skills and technology to monitor and discover illegal activities on the dark web. However, some of the law enforcement agencies still lack the technological capability to combat certain crimes [28].
- International Cooperation: Since cybercrime on the dark web is a global issue, worldwide collaboration by sharing information is required to detect it and mitigate its hazards. Some countries do have a Mutual Legal Assistance Team (MLAT)in place to allow their country's law enforcement to collaborate on that aspect. MLAT provides guidelines for sharing evidence, identifying suspects, and enforcing laws across borders. In cases where evidence is stored electronically on servers that are located in another country, MLAT helps law enforcement agencies to gain access to those servers and collect the necessary evidence [28].
- User Education: It includes training users not to download and run any random unknown software on the system, and how to spot possible malware, such as phishing emails. Campaigns and security awareness training are both necessary [28].
- Use Reputable Software: Appropriate antivirus software can identify and eliminate malware from a machine while keeping an eye on activities while it is operating. Maintaining its current status and updates with the vendor's signature is crucial [28].
- Perform Security Audits: It is critical to check a company's website often for vulnerability assessment. Early discovery of an issue may safeguard both the clients and the company [28].
- Create Regular and Verified Backups: In the event of an attack or virus, having a regular backup might make it easier to retrieve all data or other information. Ensure the security of a network: Using firewalls, IDS, IPS, and VPNs exclusively for remote access will help reduce an organization's vulnerability. Not having a reliable defense and backup can put a company in a situation where it must pay the attackers to get its system back. In 2021, Colonial Pipeline, in Texas, was targeted by ransomware that impacted their equipment managing the pipeline system. The company had to pay 75 bitcoin ($4.4 million USD) to get their system and data back [28].

Law enforcement may still use more conventional methods of fighting crime in addition to creating technology that allows them to access and deanonymize services like Tor. Some have even proposed that law enforcement may still use technological bugs or criminal errors to identify malicious actors [29].

4. METHODOLOGY

The potential of a victim of a cyber-attack or crime increases with every darknet traffic pass through the network. Detecting and blocking such traffic is considered a precautionary step to alleviate the risk of malicious activities. Generally, the process of categorizing and identifying traffic that was destined to or originated from a darknet network is known as darknet traffic classification. Because the darknet is a private network, often the traffic is encrypted and is not accessible from the regular internet unless using specific software or configurations. For that, identifying such traffic is critical to allow organizations to prevent any potential malicious activities or unauthorized behavior. Classification of the traffic involves extracting and analyzing features of the packets, such as protocols, ports, encryption, and others. With the advance of machine learning, which can utilize these features to detect darknet traffic patterns.

The proposed approach is a variation of Nearest Neighbor called LocalKNN, in which for each classified object local metric is calculated. For the sake of investigation, a comparison with other Nearest Neighbor variant algorithms was conducted to draw an inference on the performance differences in regard to darknet traffic detection. The concept of Nearest Neighbor is powerful enough that LocalKNN prioritizes local decision boundaries above global neighborhood relationships, which is a prevalent feature of darknet traffic datasets. This algorithm is particularly suitable for data with unbalanced and heterogeneous distributions. Since traditional KNNs give each neighbor equal weight and are sensitive to data distribution, this can lead to misclassification of data in complex feature spaces. On the other hand, local KNNs give closer neighbors within a specific region greater influence, enhancing the model's resistance to irrelevance and noise. In addition, reducing the effect of dimensionality through local sensitivity improves the performance of KNNs in high-dimensional spaces that are common in network traffic.

These properties make LocalKNN particularly advantageous in identifying subtle anomalies and attack patterns in darknet traffic, surpassing alternative variations such as KD-Tree or Ball Tree-based KNN in terms of detection accuracy and processing efficiency for limited feature spaces in darknet traffic patterns researchers to investigate effectiveness of different models that use the concept of extending the KNN model to classify different problems [30]. As far as our knowledge, the model used in this research has not been considered in the domain of darknet traffic classification.

Using single distance metrics can be biased toward one feature type. Therefore, this study introduces a hybrid, semantically aware distance metric for the global stage. Instead of a one-size-fits-all metric, this study proposes a combined distance function (Hybrid Global Distance Metric) that dynamically handles the mixed nature of network traffic features. The Manhattan (City-block) distance is employed for continuous features (such as packet duration, flow bytes/s). This metric is more robust to outliers than Euclidean distance in high-dimensional spaces, which is a common trait in network data. To handle the categorical features (such as protocol type, TCP flags), Hamming distance was employed, which can be optimal for comparing categorical values and measuring the number of positions at which the corresponding symbols are different. The global distance between two instances x and y is calculated as a sum of these two components:

$$Dglobal(x, y) = D_{Manhattan}(x_{cont}, y_{cont}) + D_{Hamming}(x_{cat}, y_{cat}) \quad (1)$$

This enhancement ensures that the initial neighborhood is formed based on a more accurate and meaningful measure. This step is crucial for the subsequent local metric induction to be effective. The size of the initial global neighborhood was set to 100 to ensure a sufficiently large and diverse pool of candidates for the local metric induction phase. The final local neighborhood was set to 20. This smaller value allows the algorithm to focus on the most relevant local structures without being swayed by noisy or irrelevant data points that might be included in a larger neighborhood (see Algorithm 1).

Furthermore, this study tailored the algorithm to the domain problem by using feature selection of the top 50 features via Chi-square. It is worth mentioning that using more than 50 features has a negligible gain in accuracy (< 0.3%), while reducing the feature is impacted performance by dropping accuracy (> 2%), which is likely due to the loss of key discriminative signals.

The proposed approach (Figure 1) starts with the prepossessing phase, which includes four main steps. Data cleaning, data encoding, and transforming to the appropriate form, data balancing because of the huge gap between benign and darknet data, then selecting the highest 50 rank features from the dataset. Then the resulting dataset will be used for training. The training includes training models detecting darknet traffic and another training to categorize the traffic, which helps to know what users were trying to access. Then, the trained models will be used to detect and categorize the testing dataset or any new incoming data. The details of these steps are in the following sections.

Algorithm 1. Enhanced LocalKNN prediction

Input: x (test instance), D (training data), L (training labels), k_g, k_l, cont_idx, cat_idx
1: distances ← []
2: for each d in D:
3: dist ← Σ|d[cont_idx] - x[cont_idx]| + Σ𝟙(d[cat_idx] ≠ x[cat_idx])
4: distances. append(dist)
5: global_indices ← argsort(distances)[: k_g]
6: global_labels ← L[global_indices]
7: global_dists ← distances[global_indices]
8:
9: σ ← std(global_dists)
10: weights ← exp(-(global_dists²) / (2σ²))
11: local_indices ← argsort(weights)[-k_l:]
12: local_weights ← weights[local_indices]
13: local_labels ← global_labels[local_indices]
14: return argmax_{c} Σ local_weights [local_labels == c]

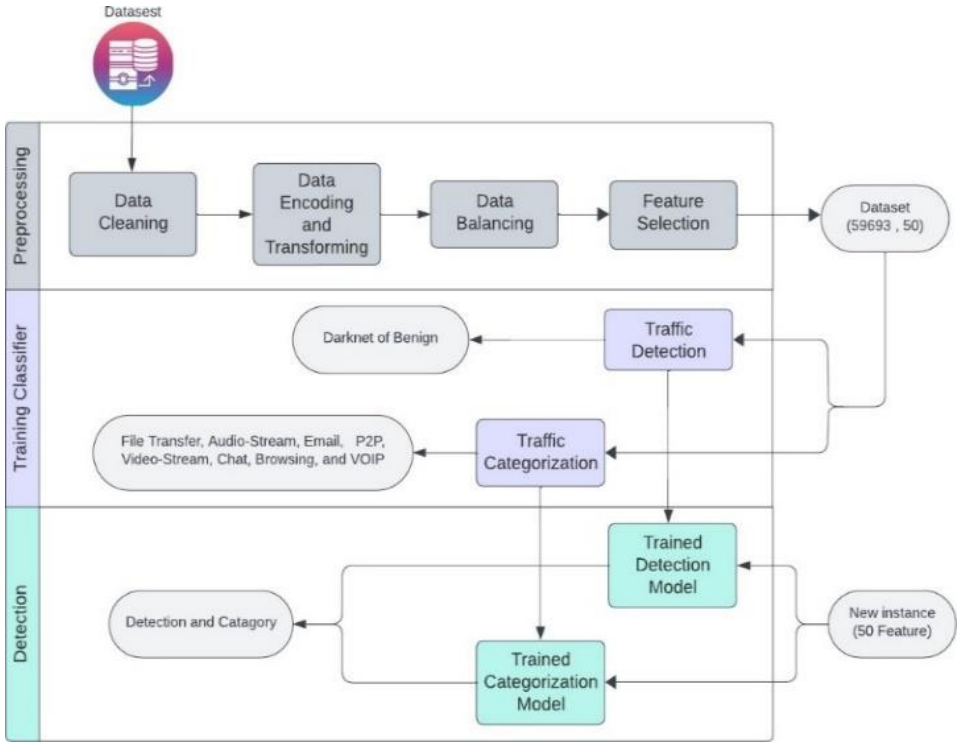

Figure 1. Diagram of the proposed approach

4.1 Prepossessing

Data preparation is a crucial phase when using machine learning techniques. It can have a significant impact on the model's performance and classification results' quality. This phase involves different sub-steps: data cleaning (handling missing values and inconsistencies), labelling, and encoding (encoding categorical data and converting timestamps), data transforming, data balancing (applying down-sampling to address class imbalance), and others. The most common and dominant steps are data missing, outliers, normalization, and encoding. These steps modified the data from raw data into more suitable data for the purpose of training a model. Overall, prepossessing can enhance the effectiveness of the model in discerning patterns in data. The following subsection will explain the preprocessing steps.

4.2 Data cleaning

The process of rectifying errors, inconsistencies, and outliers in the utilized dataset is called data cleaning. This step ensures any irregularities or inaccuracies in the data are

removed to allow the machine learning algorithm to extract relationships and patterns without being misguided by incorrect data. This includes:

- Handling missing values by removing.
- Correct inconsistent format, which mostly occurs when dealing with dates or times, which is solved by fixing the format for all datasets.
- Changing data type to match the existing data can have a significant impact on the algorithms used. The dataset has two attributes with values of float, but it was identified as string attributes, which prevents many algorithms from being useful until the format is corrected.

4.3 Data balancing

The dataset has over 141,530 traffic records, which consist of 117,219 samples of normal traffic entries and over 24,311 samples of darknet traffic entries. Even though the network traffic is not balanced in real life, but these is a huge difference in the size of the samples. This gap in the sample size could lead to a decrease in the model's effectiveness. So, balancing the data with resealable ration is needed for better performance. Therefore, this work involves using a down-sampling technique to lower the normal network traffic to 35,382 samples while keeping the number of darknet samples the same (24,311 samples). The down-sampling was chosen over oversampling (such as SMOTE) for two primary reasons: first, down-sampling reduces the total dataset size, which decreases the computational cost and training time for the KNN algorithm. Second, oversampling creates synthetic samples, which can lead to overfitting, especially on noisy data.

4.4 Data transformation

Data mostly comes in a format that may not be suitable for machine learning under investigation due to its different range of values, different data types, or other reasons. In this work, three steps were taken to transform the data into a more suitable form.

- Data Encoding: Working with network traffic data requires using IP addresses to consider originated and destined host. These IPs are in a format that reads as a string, which many learning-based algorithms cannot handle. Therefore, transforming these attributes into categorical variables is necessary. However, because the IPs can sometimes be the originating and others the destination, then the attributes of source and destination IPs must both be taken into consideration during encoding. This study uses LabelEncoding from sklearn.preprocessing package.
- Datetime Format: The datetime in the used dataset was converted into a numeric timestamp. Numeric values are more valuable in machine learning, especially when dealing with a distance metric, which involves calculating the distance between data points.
- Binary Class: The dataset has four classes, two of which include VPN and TOR network traffic, which correspond to Darknet traffic, and the other two that are non-darknet traffic. So, the classes were transformed from four classes to two classes that indicate either darknet or non-traffic.

4.5 Feature selection

Feature selection and ranking are essential steps in the prepossessing phase, which aims at reducing the dimensionality of the input data to decrease the computational cost of later processes and focus attention on the most relevant data points. This step involves selecting a subset of relevant and informative features while eliminating noisy and redundant features that could have little unnoticeable impact on the ability of the model to classify the target class. Furthermore, feature selection can mitigate the over-fitting issue by using lower feature dimensions to generalize the model. This work proposes using a widely utilized filter type selection algorithm called Chi-Square. Chi-Square is a statistical test to select a subset of features through examining the relationship between the features by assessing the correlation strength of features using computed statistical values (Chi-Square value). This value is calculated between each feature i and the corresponding target class to examine the correlation. The feature will be discarded if there is no correlation, feature and class are independent. Otherwise, the feature will be selected to be impotent because there is a dependency between the feature and the target class. Based on the importance, features are ranked from highest to lowest Chi-Square value, and then the desired number of features is selected. This research seeks to select 50 features from 82 features in the dataset [31].

4.6 Classification

This research proposed a model that depends on one of the Nearest Neighbor variation algorithms, which is called LocalKNN. LocalKNN is an extension of KNN, which involves calculating a local metric as an extra step for each classified object. During the process of classifying test objects, classifiers use global metrics to find a large set of nearest neighbors, which is then used to derive a new local metric set. Finally, a locally induced metric is used to select the nearest neighbors. This method improves classification accuracy, particularly for the case of data with nominal attributes, and is also reasonable to use this method for large data sets [32].

LocalKNN, like any Nearest Neighbor algorithm, used distance metric to calculate the distance between objects. The distance metric used in this model is CityAndHamming, which is a combination of the city block Manhattan metric and the Hamming metric. In Manhattan metrics, the distance is calculated by taking the sum of absolute differences between values across all the dimensions (x and y). The Hamming distance is calculated by measuring the similarity of two numbers by considering each corresponding digit position that is different, which means 0 for equal and 1 for different. These metrics are used from the library proposed by Bazan and Szczuka [32]. The concept of Nearest Neighbor is an attentive topic that can provide a model with fast and accurate comparing other algorithms. Investigating the strength of the multiple approaches can provide an understanding of the general aspect of using Nearest Neighbor algorithms in the dark net domain. To do that, this paper ran different experiments using the model listed below to draw the inference. This section reviews the seven other nearest-neighbor-based classification algorithms that were applied in this work:

- KNN: This algorithm was used along with the Manhattan distance to provide high performance

model. It aims to find the KNN point for the current point. The class prediction depends on the majority votes of the neighbors according to the Manhattan metric. The performance of the KNN model depends on two factors: the distance metric and the value of k (the number of neighbors) [33].

- KStar: K-Star or K* algorithm is an instance-based classifier that uses the KNN method. It aims to divide data into k clusters. K* as a classifier uses an entropic distance measurement depending on the probability of transforming one instance into another and calculating the inter-instance distance of training data samples. K* can provide high performance and strong generalization, especially with balanced data. K* can be a valuable strategy for handling the real-valued and symbolic at tribute with great way in dealing with missing values. However, it can require longer training time [34].
- Extended Nearest Neighbor (ENN): It is a kind of an advanced version of the KNN classifier. In classic KNN, estimating group membership only considers the nearest neighbors of the sample. In contrast, ENN was extended to go one step further by considering samples that consider the test sample as their nearest neighbors. In this way, ENN can maximize the intra-class coherence gain over the whole data set. However, the ENN is computationally more expensive than KNN, especially with a high-dimensional dataset [35].
- Nearest Centroid: In this classifier, the centroid for each class is calculated, which later can be used to predict the class of any encountered data point example by finding the nearest centroid using the provided distance. Hamming distance was used with this classifier to calculate the nearest centroid [36].
- Nearest neighbor-like algorithm using Non-Nested Generalized (NNge): this classifier combines instance-based learning with rule induction to obtain the advantages of both paradigms. In instance-based learning (lazy), most computation is performed during the prediction phase, which allows new instances to be included over time, and this can be effective immediately, even with small input data. Due to that, the amount of the stored instance can increase significantly. In rule induction (eager), computation is performed during the learning phase, which results in a small set of less expensive computational rules. These rules, or generalized rules if instance generalized into rules (called exemplars), can be stored for prediction purposes. Generalized rules must not overlap, which enforces the non-nested aspect of the approach. The algorithm joins a new instance into the nearest generalized exemplars, using a distance metric that has the same severity. This approach aims to balance the advantages of instance-based learning and rule induction while ensuring efficient and non-overlapping generalization [37].
- ReslibKNN: It combines KNN with rule induction. We propose an algorithm that preserves lazy learning; rules are reconstructed in a lazy way at the moment of classification, like the nearest neighbors. The proposed combination uses the metric-based generalization of rules. This is similar to NNge, except that NNge is not a lazy technique and does not allow overlapping. For the distance metrics, City and Hamming were used [32, 38].
- KNN using Log and Gaussian weight kernels (KNNLG): It is KNN based approach that uses a weighted distance, which can be the similarity of the inverse of the distance. Similarly, it can be associated with weight calculation using the negative logarithm or a Gaussian. A Gaussian is assumed for every KNN, and weights are associated relative to the distance of the neighbor from the mean in the Gaussian. In this work, the Gaussian is used as a method to calculate the Weight for each distance, which is calculated using the City-block Manhattan metric [39].

5. EXPERIMENTAL RESULTS

The proposed model went through an evaluation phase that aimed to judge the performance in the domain of darknet traffic detection and categorization. This section presents the findings of the conducted experiments that were performed to evaluate different nearest neighbor variation classification algorithms. The focus is on providing a model that can highly detect darknet traffic, along with determining the type of traffic. The experiment has different steps that were explained in the previous section. Steps include data pre-processing, encoding, balancing, transforming, and feature selection using Chi-Square, then the resulting dataset is fed into eight classification algorithms that depend on or use the concept of nearest neighbor. The models were implemented in Python utilizing Scikit-Learn and WEKA that is a widely used open-source toolkit for machine learning experiments.

5.1 Dataset description

This paper utilized a dataset, called CICDarknet2020, from the Canadian Institute for Cybersecurity. The dataset was produced using dual layered approach to obtain the data. The first layer was used to produce benign or darknet traffic. The second traffic was to generate the categories of the traffic, which include File Transfer, Audio-Stream, Email, P2P, Video-Stream, Chat, Browsing, and VOIP. This dataset was generated by integrating two previously generated datasets called ISCXTor2016 and ISCXVPN2016. In this dataset, VPN and TOR correspond to darknet traffic. Table 2 shows the information on the total record of each class in the dataset before and after balancing. The darknet traffic was generated for eight diverse types of categories, as shown in Figure 2, with the total number of records for each one.

5.2 Darknet traffic detection

This section discusses the performance of classification methods used during the experiments to identify darknet traffic from benign traffic. This algorithm was evaluated on a darknet traffic dataset that went through reprocessing. As part of the reprocessing, a Chi-Square features selection approach was used to select the top 50 features that are most relevant to the target class. The classifiers provide accurate results, which ranged from 86% to 99.34%. The lowest performance was obtained using the Nearest Centroid approach, which was able to classify the traffic correctly only 98% of the time. This is considered very low in the domain applied. KStar approach does not show good potential in determining the traffic; it obtains an accuracy of 82.61%. That makes KStar and Nearest

Centroid not an option in the domain of detecting darknet traffic. The other algorithms were able to perform up to the expectation and obtain high accuracy detection, with results ranging from 97%–99.34%. The Performance of LocalKNN was the highest obtained from the experiments, with an accuracy of 99.34%, which gives high potential to be investigated further. ReselibKNN was the runner-up with an accuracy of 99.19%, which is comparable to LocalKNN. These two approaches were able to perform higher than the standard KNN approach, that was obtain accuracy of 98.67%. Their high performance can be due to their ability to draw a generalization using the example encountered to classify any new example, and their approach of including more instances in the process of finding the nearest neighbors. Table 3 shows the accuracy of the classifiers used during the experiments.

Table 2. Dataset classes total records

Class	Class Type	Records	Total	After Balancing
Benign	Non-TOR	93,356	117,219	35,382
	Non-VPN	23,863		
Darknet	VPN	22,919	24,311	24311
	TOR	1392		

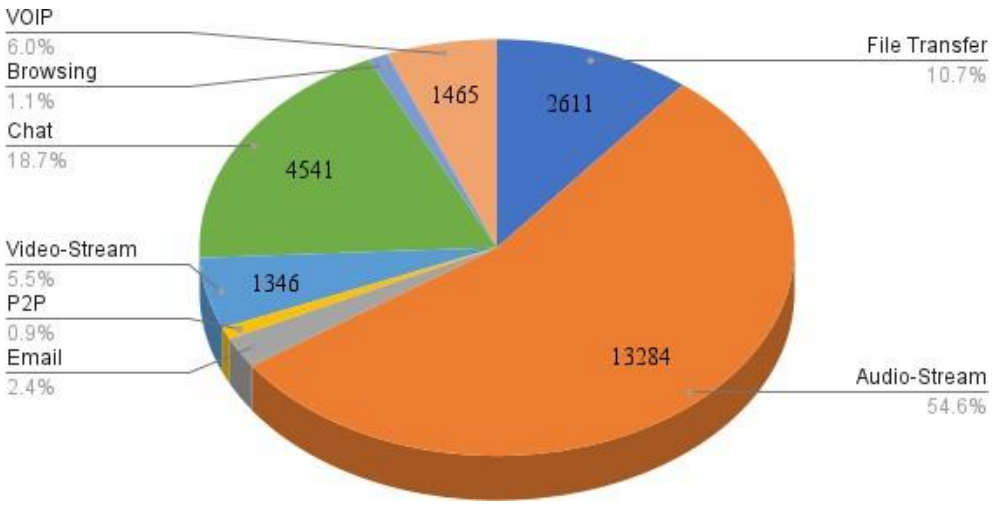

Figure 2. Darknet traffic categories

Table 3. Darknet traffic classification accuracy

KNN	RseslibKNN	KStar	ENN	Nearest Centroid	NNge	KNNLG	LoaclKNN
98.67	99.19	82.61	97.00	68.00	98.51	98.68	99.34

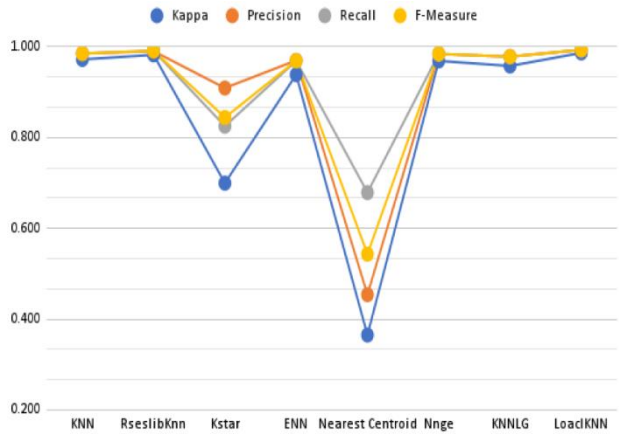

Figure 3. Evaluation metrics results

To confirm the findings from the experiments, Precision, Recall, F-Measure, and Kappa were calculated as part of the approach's evaluations. These metric results were aligned with the accuracy results to confirm that LocalKNN and ReselibKNN were better approaches in tackling the darknet traffic detection than other approaches. They both have a value of over 99% for Precision, Recall, F-Measure, and Kappa over 98%. Notably, RseslibKnn and LoaclKNN were able to achieve the highest Kappa, that indicating strong agreement between actual and predicted classes. In general, LocalKNN demonstrates the highest results across all metrics – precision 99.3%, recall 99.3%, F-Measure 99.3%, and Kappa 98.86%– which highlights its effectiveness in darknet classification.

At the same time, Nearest Centroid and KStar had the lowest values across all metrics, which indicates they are not effective for the darknet traffic detection problem. Figure 3 illustrates the results of the evaluation metrics.

5.3 Darknet traffic category classification

Detecting the darknet traffic can help alleviate the risk that is associated with accessing it. However, determining the category of the traffic can help understand the users' needs and behaviors in terms of why they are accessing the darknet in the first place. To evaluate the ability of the models to distinguish the type of traffic, categorization results were collected. In this experiment, the five best performance models were considered; therefore, ENN, KStar, and Nearest Centorid approach were eliminated as they performed poorly and did not improve the standard KNN approach.

Categorizing the darknet traffic showed similar results

where LocalKNN provided the best performance with an accuracy of 91.34% comparing to other models. RselibKNN was the second in the list with an accuracy of 89.77%. Table 4 shows the results of the models.

The evaluation metrics were also calculated to validate the findings, which emphasize that using LocalKNN proven to be better at dealing with darknet traffic categorization as well as detection. Figure 4 illustrates the metrics results.

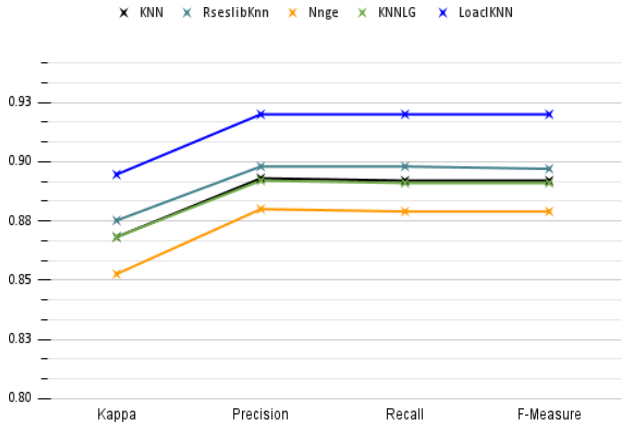

Figure 4. Evaluation metrics results for darknet traffic categorization

Table 4. Darknet traffic categorization accuracy

KNN	RseslibKNN	NNge	KNNLG	LoaclKNN
89.11	89.76	87.89	89.11	91.34

Table 5. Evaluation metrics for categories of darknet traffic

Class	Precision	Recall	F-Measure
Audio-Streaming	0.95	0.94	0.94
Browsing	0.91	0.95	0.93
Chat	0.88	0.86	0.87
Email	0.79	0.75	0.77
File-Transfer	0.87	0.84	0.85
P2P	0.99	0.99	0.99
Video-Streaming	0.71	0.74	0.72
VOIP	0.84	0.87	0.86

The metric evaluation for each class in the categorization in Table 5. The model was able to achieve a high F-measure on P2P, Audio-Streaming, and browsing, ranging from 0.93 to 0.99. However, the model performs less accurately when dealing with Video-Streaming and Email with F-Measure of 0.72 and 0.77, respectively. The rest of the class is falling in between. This means the traffic of P2P is easier to recognize than any other traffic class.

6. CONCLUSIONS

The dark web's anonymity makes it a fertile ground for crimes due to the difficulty of tracking or revealing the identities of users. In this paper, the dark web was reviewed to provide a better understanding of the activities that occur in it, along with the risks associated with accessing its sites. The usage of "dark" has its own implications and impact on society and law, which were discussed. Therefore, this paper addressed the challenges of dark net traffic by utilizing various classification models that depend on or use the nearest neighbor concept, including (KNN, Reslib KNN, KStar, ENN, Nearest Centroid, Nearest neighbor like algorithm using NNge, KNNLG, and LocalKNN). The proposed system was evaluated using five metrics: accuracy, precision, recall, F-measure, and Kappa. Through experiments on the CICDarknet2020 dataset. The proposed approach suggests using LocalKNN, which obtains the highest accuracy of 99.34% in detecting traffic and 91.34% for categorizing the traffic. The model used a prepossessing phase with selecting 50 features using the Chi-Square approach. At the same time, the results show that using Nearest Centroid and KStar are not viable choices when it comes to dealing with darknet traffic data because of their deficient performance, with an accuracy of 68% and 82%. Future directions for this research include applying the algorithm to different traffic datasets on the dark web and in the real world, investigating hybrid models that combine this algorithm with other deep learning techniques in order to enhance detection capabilities, and then comparing the results with those achieved in this research.

REFERENCES

[1] Al-Omari, A., Allhusen, A., Wahbeh, A., Al-Ramahi, M., Alsmadi, I. (2022). Dark web analytics: A comparative study of feature selection and prediction algorithms. In 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), San Antonio, TX, USA, pp. 170-175. https://doi.org/10.1109/IDSTA55301.2022.9923042

[2] Parkar, A., Sharma, S., Yadav, S. (2017). Introduction to deep web. International Research Journal of Engineering and Technology (IRJET), 4(6): 229-234. https://www.irjet.net/archives/V4/i6/IRJET-V4I6515.pdf.

[3] Besenyő, J., Gulyas, A. (2021). The effect of the dark web on the security. Journal of Security & Sustainability Issues, 11(1): 103-121. https://doi.org/10.47459/jssi.2021.11.7

[4] Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., Lenders, V. (2019). BlackWidow: Monitoring the dark web for cyber security information. In 2019 11th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, pp. 1-21. https://doi.org/10.23919/CYCON.2019.8756845

[5] Han, C., Shimamura, J., Takahashi, T., Inoue, D., Kawakita, M., Takeuchi, J.I., Nakao, K. (2019). Real-time detection of malware activities by analyzing darknet traffic using graphical lasso. In 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, pp. 144-151. https://doi.org/10.1109/TrustCom/BigDataSE.2019.00028

[6] Habibi Lashkari, A., Kaur, G., Rahali, A. (2020). Didarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning. In Proceedings of the 2020 10th International Conference on Communication and Network Security, Tokyo, Japan, pp. 1-13. https://doi.org/10.1145/3442520.3442521

[7] Mohanty, H., Roudsari, A.H., Lashkari, A.H. (2022). Robust stacking ensemble model for darknet traffic

classification under adversarial settings. Computers & Security, 120: 102830. https://doi.org/10.1016/j.cose.2022.102830
[8] Almomani, A. (2025). Darknet traffic analysis, and classification system based on modified stacking ensemble learning algorithms. Information Systems and e-Business Management, 23(1): 209-240. https://doi.org/10.1007/s10257-023-00626-2
[9] Yadav, D., Bhushan, B., Saxena, S. (2020). The dark web: A dive into the darkest side of the Internet. SSRN Electronic Journal, 10. http:// doi.org/10.2139/ssrn.3598902
[10] Sobhan, S., Williams, T., Faruk, M.J.H., Rodriguez, J., et al. (2022). A review of dark web: Trends and future directions. In 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), Los Alamitos, CA, USA, pp. 1780-1785. https://doi.org/10.1109/COMPSAC54236.2022.00283
[11] Alshammery, M.K., Aljuboori, A.F. (2022). Crawling and mining the dark web: A survey on existing and new approaches. Iraqi Journal of Science, 63(3): 1339-1348. https://doi.org/10.24996/ijs.2022.63.3.36
[12] Kaur, S., Randhawa, S. (2020). Dark web: A web of crimes. Wireless Personal Communications, 112(4): 2131-2158. https://doi.org/10.1007/s11277-020-07143-2
[13] Bhushan, B., Saxena, S. (2020). The dark web: A dive into the darkest side of the internet. In International Conference on Innovative Computing and Communication (ICICC 2020).
[14] Jardine, E. (2015). The dark web Dilemma: Tor, anonymity and online policing. Global Commission on Internet Governance.
[15] Bhushan, B., Saxena, S. (2020). The dark web: A dive into the darkest side of the internet. In International Conference on Innovative Computing & Communication (ICICC) 2020.
[16] Gupta, A., Maynard, S.B., Ahmad, A. (2021). The dark web phenomenon: A review and research agenda. ACIS. https://api.semanticscholar.org/CorpusID:150024410.
[17] Susuri, A. (2019). Dark web and its impact in online anonymity and privacy: A critical analysis and review. Journal of Computer and Communications, 7(3): 30-43. https://api. semanticscholar.org/CorpusID:108530073.
[18] Dawood, O.A., Rahma, A.M.S., Hossen, A.M.J.A. (2017). New symmetric cipher fast algorithm of revertible operations' queen (FAROQ) cipher. International Journal of Computer Network and Information Security, 9(4): 29-36. https://doi.org/10.5815/ijcnis.2017.04.04
[19] Dawood, O.A., Rahma, A.M.S., Hossen, A.M.J.A. (2015). The new block cipher design (Tigris Cipher). International Journal of Computer Network and Information Security, 7(12): 10-18. https://doi.org/10.5815/ijcnis.2015.12.02
[20] Nazah, S., Huda, S., Abawajy, J., Hassan, M.M. (2020). Evolution of dark web threat analysis and detection: A systematic approach. IEEE Access, 8: 171796-171819. https://doi.org/10.1109/ACCESS.2020.3024198
[21] Upulie, H.D.I., Prasanga, P.D.T. (2021). Dark web, its impact on the internet and the society: A review.
[22] Nazah, S., Huda, S., Abawajy, J., Hassan, M.M. (2020). Evolution of dark web threat analysis and detection: A systematic approach. IEEE Access, 8: 171796-171819. https://doi.org/10.1109/ACCESS.2020.3024198
[23] Basheer, R., Alkhatib, B. (2021). Threats from the dark: A review over dark web investigation research for cyber threat intelligence. Journal of Computer Networks and Communications, 2021(1): 1302999. https://doi.org/10.1155/2021/1302999
[24] Chertoff, M., Simon, T. (2015). The impact of the dark web on internet governance and cyber security. Global Commission on Internet Governance, 6. https://www.cigionline.org/sites/default/files/gcig_paper _no6.pdf.
[25] Han, C., Takeuchi, J.I., Takahashi, T., Inoue, D. (2021). Automated detection of malware activities using nonnegative matrix factorization. In 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China, pp. 548-556. https://doi.org/10.1109/TrustCom53373.2021.00085
[26] Mahmood, I., Rahman, M.A., Kabir, M.A., Shahriar, M., Rahman, P. (2022). A survey on dark web monitoring and corresponding threat detection.
[27] Salman, A.D., Hasan, E.H. (2023). Survey study of digital forensics: Challenges, applications and tools. In 2023 16th International Conference on Developments in eSystems Engineering (DeSE), Istanbul, Turkiye, pp. 788-793. https://doi.org/10.1109/DeSE60595.2023.10469020
[28] Cole, R., Latif, S., Chowdhury, M.M. (2021). Dark web: A facilitator of crime. In 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Mauritius,
[29] Naseem, I., Kashyap, A.K., Mandloi, D. (2016). Exploring anonymous depths of invisible web and the digi-underworld. International Journal of Computer Applications, 3: 21-24. https://pdfs.semanticscholar.org/47d6/8ea79f06ab80f15 69aabefd1bc74ca30a541.pdf.
[30] Uddin, S., Haque, I., Lu, H., Moni, M.A., Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1): 625. https://doi.org/10.1038/s41598-022-10358-x
[31] Zhai, Y., Song, W., Liu, X., Liu, L., Zhao, X. (2018). A chi-square statistics based feature selection method in text classification. In 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, pp. 160-163. https://doi.org/10.1109/ICSESS.2018.8663882
[32] Bazan, J.G., Szczuka, M. (2000). RSES and RSESlib-a collection of tools for rough set computations. In International Conference on Rough Sets and Current Trends in Computing, Banff, Canada, pp. 106-113. https://doi.org/10.1007/3-540-45554-X_12
[33] AlZoman, R.M., Alenazi, M.J. (2021). A comparative study of traffic classification techniques for smart city networks. Sensors, 21(14): 4677. https://doi.org/10.3390/s21144677
[34] Ghasemkhani, B., Aktas, O., Birant, D. (2023). Balanced k-star: An explainable machine learning method for internet-of-things-enabled predictive maintenance in manufacturing. Machines, 11(3): 322. https://doi.org/10.3390/machines11030322
[35] Tang, B., He, H. (2015). ENN: Extended nearest neighbor method for pattern recognition [research frontier]. IEEE Computational Intelligence Magazine,

10(3): 52-60. https://doi.org/10.1109/MCI.2015.2437512
[36] Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences, 99(10): 6567-6572. https://doi.org/10.1073/pnas.082099299
[37] Bohacik, J., Zabovsky, M. (2017). Nearest neighbor method using non-nested generalized exemplars in breast cancer diagnosis. In 2017 IEEE 14th International Scientific Conference on Informatics, Poprad, Slovakia, pp. 40-44. https://doi.org/10.1109/INFORMATICS.2017.8327219
[38] Wojna, A. (2005). Analogy-based reasoning in classifier construction. In Transactions on Rough Sets IV, pp. 277-374. https://doi.org/10.1007/11574798_11
[39] Sreenivasamurthy, S., Frank, S. (2015). Efficacy of season prediction for geo-locations using geo-tagged images. In 2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, pp. 476-484. https://doi.org/10.1109/IRI.2015.79