International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Different Machine Learning Algorithms used for Secure Software Advance using Software Repositories

Kanchan Chaudhary, Dr. Shashank Singh
Department of Computer science and Engineering, Integral University, Lucknow, Uttar Pradesh, India

ARTICLEINFO

Article History:

Accepted: 01 April 2023
Published: 11 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
300-317

ABSTRACT

In the present phase of the Fourth Industrial Revolution (4IR or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding smart and automated applications, the knowledge of artificial intelligence (AI), particularly, machine learning (ML) is the key. Cyber Security attacks are significantly growing in today's modern world of technology and advanced software development. The inclusion of cyber security defense is vital in every phase of software development. Identifying and implementing key relevant cyber security vulnerability controls during the early stages of the software development life cycle, i.e., the requirement phase is especially important. The Common Attack Pattern Enumeration & Classification (CAPEC) is a publicly available software repository from MITRE that currently lists 555 vulnerability attack patterns. As Cyber Security continues to exponentially grow in complexity, the importance of the Machine Learning role to automate the identification of vulnerabilities for various software development is paramount to aid software developers in creating protected software. This paper discusses the conducted survey on different machine learning algorithms used for secure software development using software repositories.

Keywords: Machine Learning, Topic Modeling, Cyber Security, CAPEC, MITRE

I. INTRODUCTION

Cyber security is the field of implementing methods of protecting networks, devices and data from unauthorized access and the practice of protecting confidentiality, integrity, and availability of digital information according to CISA, 2019. There has been a 600% increase in cybercrime since January of 2020. Small businesses have seen 43% of these attacks with an increase of 400% during this period.

Cyber security attacks are estimated to grow to a staggering $10.5 trillion in cost of damages by 2030. Vulnerabilities and software programming errors are being exploited by attackers through flaws in software, firmware, and hardware. Software repositories identify, define, and catalog publicly disclosed cybersecurity vulnerabilities. There are several publicly available resources available including the Common Attack Pattern Enumeration & Classification (CAPEC), Common Vulnerabilities and Exposures (CVE), Common Weakness Enumeration (CWE), and National Institute of Standards and Technology (NIST) Cyber security Framework. As these repositories are updated, software needs to be updated to include protection from the newly identified threats. During the requirements phase, the early stages of the software development life cycle, identifying key relevant cyber security vulnerability controls will drastically save time, money, and increase protection from vulnerabilities. Implementation during initial coding will reduce the amount of time recoding to satisfy vulnerability protection. As Cyber Security continues to exponentially grow in complexity, the importance of the Machine Learning (ML) role to automate the identification of vulnerabilities for various software development is paramount to aid software developers and companies in creating safe and secure software during the requirements phase. Automation of the identification of most significant vulnerabilities to planned software would reduce human error and labor intensity resulting in financial savings and protection of sensitive data.

This research discusses ML algorithms and their potential use for secure software development through implementation of software repositories. This paper is organized as follows: Section 2 discusses background and literature regarding software repositories and implementation in computer programming. Section 3 discusses literature related to cyber security and ML. Section 4 discusses several ML algorithms and their current or potential uses. Section 5 discusses further evaluation of software repositories and related research. Section 6 draws conclusion of the research and the development of this research.

II. METHODS AND MATERIAL

Software repositories are vital in providing information regarding vulnerabilities and solutions to improve cyber defense. These repositories commonly overlap information and store the information in different formats. Many companies and government agencies maintain their own software repository which are developed from public repositories such as CAPEC or The National Institute of Standards and Technology (NIST). MITRE is a nonprofit organization that serves public interest as an independent advisor for advancement in technology and national security. MITRE is a federally funded organization that is currently working with US Government sectors such as the Veterans Affairs (electronic-medical-record standards project) and Department of Homeland Security (fingerprint identification project) according to Macsai (2012). MITRE is currently facilitating over two hundred independent projects as well as CAPEC, CVE, CWE and MITRE ATT&CK.Common Vulnerabilities and Exposures (CVE) identify, define, and catalog publicly disclosed cybersecurity vulnerabilities. CVE is sponsored by the U.S. Department of Homeland Security (DHS) and Cybersecurity and Infrastructure Security Agency (CISA) as stated on cve.org. There are currently 187,544 CVE records accessible to the public. The CVE repository is a list of vulnerability events (records) that are reviewed and certified by the CVE Numbering Authority (CNA). A CVE entry includes detailed attack incident information categories include Name, Status, Description, References, Phase, Votes, and Comments. The repository is available for download in various forms such as CSV, HTML, Text, XML and JSON 4.0. These certified vulnerabilities are specific events useful for research in comparing individual threat types. This data is used to extrapolate statistics by organizations such as MITRE. The Common Attack Pattern Enumeration & Classification (CAPEC) classifies and organizes cyber security attack patterns for public use. CAPEC defines attack patterns used by cyber criminals to infiltrate software through new or established vulnerabilities. This cyber security repository is used by analysts, developers, testers, and educators to advance community understanding and enhance defenses (MITRE, 2022). IBM and Computer Aided Integration of Requirements and Information Security (CIARIS) are two prime examples of organizations that utilize CAPEC as a foundation to software security (MITRE, 2022).There are several software options available that scan programs for vulnerabilities based off of CAPEC or a combination of other repositories. CAPEC entries can be downloaded in HTML, CSV, or XML formats. CAPEC has several organized options such as CAPEC Navigation, External Mappings, and Helpful Views to download and view the entries. A Navigation view of Mechanism of Attack entry includes Attack Pattern ID, Name, Abstraction Type, Status, Description, Alternate Terms, Likelihood of Attack, Typical Severity, Related Attack Patterns, Skills Required, Resources Required, Indicator, Consequences, Mitigations, Example Instances, Related Weaknesses, Taxonomy Mappings, and Notes. Common Weakness Enumeration (CWE) is a community developed repository listing weakness types for software and hardware (Kanakogiet al., 2022). The searchable CWE repository is available and free to reference over six hundred weaknesses. Several companies such as Apple, HP, IBM, and Red Hat are contributing to the development of CWE entries. The CWE list is currently offered in many views including Dictionary, Classification Tree, Graphical and Slices by Topic. CWE is referenced or implemented by individuals, companies and the U.S. National Vulnerability Database (NVD). CWE is commonly paired with CAPEC for a greater understanding of software security defense.CWE can be downloaded as a HTML, CSV or XML file other options include PDF and XSD. Navigate CWE, External Mappings, Helpful Views and Obsolete Views are Categories of download views. A CWE entry for an external Mappings CWE, Top 25 (2022) includes CWE-ID, Name, Weakness Abstraction, Status, Description, Extended Description, Related Weakness, Weakness Ordinalities, Application Platforms, Background Details, Alternate Terms, Modes of Introduction, Exploitation Factors, Likelihood of Exploitation, Common Consequences, Detection Methods, Potential Mitigations, Observed Examples, Related Attack Patterns and Notes. MITRE also offers the Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) repository. ATT&CK is an attack matrix comprised of fourteen categories of tactics. These tactic categories contain definitions of attack types, mitigation strategy, detection strategy and additional sub-techniques pertinent to the technique to counteract attacks. The ATT&CK knowledge database includes profiles of large cyber offending groups, which is substantially helpful in the event of an attack occurring. The attacker's profile provides details on their past tendencies and tactics. This model is particularly helpful if you have identified a specific vulnerability by providing detailed information about the vulnerability and provided methods to prevent exploitation of that weakness.The ATT&CK Matrix for Enterprise has fourteen Tactic categories with the number of techniques listed: Reconnaissance (10), Resource Development (7), Initial Access (9), Execution (13), Persistence (19), Privilege Escalation (13), Defense Evasion (42), Credentials Access (17), Discovery (30), Lateral Movement (9), Collection (17), Command and Control (16), Exfiltration (9), and Impact (13) (MITRE ATT&CK, 2022). Each Tactic entry lists the ID, definition, and associated techniques. Each technique has an ID, definition, list of sub techniques, Mitigations and Detection sections. The MITRE ATT&CK website goes in depth defining software used in attacks and description of threat groups and the tactics and techniques used. MITRE also offers Matrix variations for Enterprise such as: PRE, Windows, MacOS, Linux, Cloud, Network, and Containers. A matrix for Mobile platforms: Android and iOS is available along with the Matrix for ICS (Industrial Control Systems). The National Institute of Standards and Technology (NIST) is a government agency that develops technology, metrics, and standards for innovation and economic competitiveness for United States based organizations (NIST, 2019). The guidelines recommended by NIST are sets of security controls used for information systems which have been endorsed by the government and include the best security practices for a range of industries. NIST provides guidelines for information systems these recommendations have been endorsed by the government and include the best security practices for a range of industries. One example of NIST standards includes the NIST Cybersecurity Framework. This framework helps to improve the risk management for businesses which have adopted this model and although it does not fit the needs of all organizations and businesses, it has succeeded in reducing and mitigating many risks that organizations come across. NIST offers the National Vulnerability Database (NVD).

III. LITERATURE SURVEY

This research serves as continuing research and education for development of an automated recommender system for identifying top correlated CAPEC attack patterns found from a Software Requirements Specifications (SRS) document (Vanamalaet al., 2020) . Previous research leading to this project include “Topic Modeling and Classification of CVE Database” (Vanamalaet al., 2019) and “Analyzing CVE Database Using Unsupervised Topic Modelling” (Vanamalaet al., 2020).

Previous research of CVEs from the National Vulnerability Database was conducted in 2019 by use of Topic Modeling (Vanamala et al., 2019). The topic model used was Gensim’s Latent Direct Allocation (LDA) and visualized the topics with pyLDAvis. The process included: Data Processing, Text Cleaning, Bigram and Trigram Models, Building the Topic Model, Visualization of the Topics and the Associated Keywords, and the final phase, Mapping Topics to Open Web Application Security Project (OWASP)-Top 10. OWASP is a non-profit organization providing practical information regarding application security which is used by security professionals and developers. The CVE dataset used included 121,716 CVE entries from 1999 to 2019. The topic model was used to identify trends in 5 year increments. These trends were then manually compared to the OWASP Top-10 list relative to the 5 year increments. The results of the topic model revealed large similarity to the OWASP Top-10 recommendations provided for the same time frames. This revealed the accuracy of the topic model in comparison to the OWASP expert analysis.

“Topic Modeling and Classification of CVE Database” was conducted (Vanamala et al., 2020). This research expanded on the previous project and additionally implemented the use of LDA ML for OWASP Top-10. The LDA identified correlated topics for the CVEs and the OWASP Top-10. Results of both lists would now be compared through the implementation of statistical analysis by use of Standard deviation and Coefficient of Variance. The results comparing the first project (manually mapping) and this project (automated mapping) resulted in remarkably similar values. These results indicated that the automated mapping process was able to closely replicate the manual mapping process from the first project. The development of the automated mapping comparison between two large data sets was concluded as a success. This ML model was able to reproduce expert analysis.

In recognition of CAPECs large database, “Recommending Attack Patterns for Software Requirements Document” (Vanamala et al., 2020) was developed. The transition to CAPEC from CVE data was implemented in efforts to provide more practical assistance to Software developers during the Requirements Engineering (RE) phase. Topic models for the SRS document and CAPEC were developed for the same automated mapping style in the previous project. A Software Requirements Specification (SRS) document outlines the desired capabilities and results that the proposed program should encompass. Identifying and implementing cyber security vulnerability controls directly from the SRS document significantly improves the inclusion of cyber security protection during the development of the software. Manually evaluating and finding the most relevant CAPEC attack patterns from the CAPEC list of 555 attack patterns is time consuming and requires an elevated level of expertise. The automation of this process of ML through use of LDA alleviates the human error and time intensity of that objective. Cosine Similarity was the metric process used to identify the similarity in topic distribution between keywords identified from CAPEC and the SRS Document. The advantage of using Cosine Similarity incorporates not only the measure of how similar the documents are irrespective of their sizes, mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The topic model was successfully able to identify relevant CAPEC attack patterns from the SRS document.

Review of recent and related research was evaluated including research topics, cyber security and/or ML research. The research papers were reviewed for type of ML, metrics used and/or any relevancy directly to cyber security. The goal is to identify any new process or information useful in further developing our current research.

Research focused on making connections between CVE and CAPEC was conducted (Kanakogi et al., 2022). The research goal was to achieve a more comprehensive cyber security repository by creating a link between the two databases automatically using Artificial Intelligence(AI). CVE provides information on specific events that are detailed and CAPEC provides a classification of attack patterns. Both are useful on their own, a database linking information from both would significantly increase user full understanding of the attack type with correlated examples of attacks. The researchers linked CAPEC-IDs to the CVE-IDs implementing and comparing results of three algorithms: SBERT, TF-IDF and USE. Each algorithm was used with three metrics/experiment patterns: Document, Per Section and Section Average. The exception was for SBERT and Document due to token limitations. The research concluded that TF-IDF was the most suitable overall for this testing. The researchers suggest further research to utilize Ensemble Learning which combines results from different tests to achieve an average score. Combining multiple professional sources for cyber security vulnerabilities would be invaluable to Cyber Security Experts and programmers across the world. This research coincides with our current research in terms of a goal of assisting programmers in connecting cyber security information to users through an automated process.

Research focused on creating a template for evaluating security status for established networks aimed to provide a checklist to assist Chief Security Officers to identify insufficient security tasks (Alyami et al., 2021). This checklist addresses thirty questions related to: Operating systems, IPS software, Antivirus Software, Anti-Bot software, URL software, Application Control Software, Zero-day protection and DDoS solutions. Additionally, there is a methodology or strategy table to assist or direct on how to evaluate and how to mitigate risks if found. This is a research topic specific to cyber security threats in the form of assisting in evaluating current security measures implemented. Many cyber security related research projects incorporate the evaluation of current protection status. This is a significantly useful tool for ongoing upkeep of established programming and networks. Our research aims to help programmers during initial software creation phase to allow for safe and secure programming before implementation of the program.

The next evaluated research was on the creation of the Security Assessment Model (SAM). Guru Prasad et. al., (2022) explained that SAM is a complex multifaceted model that evaluates software for security risk. The authors focused on utilizing well established resources such as CWE, ISO/IEC 25010, SMARTS/SMARTERS and AHP. The weighted system was evaluated on 150 popular open source JAVA applications through GitHub along with 1200 test cases from OWASP Benchmark. This large data set was critical to validate and assess the operational capability of SAM. The outcome provided a success as it was able to classify the level of security but also produce a weighted list for developers to tackle. The weighted list eliminates the triage process, saving time. At the time of the research conclusion, this was the only model that went to the extent of properly analyzing and producing a working priority list of deficiencies for Java programs.

Machine Learning

Machine Learning, a form of Artificial Intelligence (AI), is used to accomplish tasks involving large data and automated detection of meaningful patterns in data (Shalev-Shwartz and Ben-David, 2014). ML is used to solve tasks too complex or too time consuming for humans to complete. There are two main categories of ML: supervised and unsupervised. The significant difference between these two types of ML is the use of labeled or unlabeled data. Supervised ML uses a labeled data set as a training set. According to Shalev-Shwartz and Ben-David (2014) advanced algorithms then compare unlabeled data to the labeled data set to derive correlation of topics or discover hidden meaning to the text. Unsupervised ML does not have a distinction between training and test data used.

Unsupervised Machine Learning

Unsupervised learning utilizes ML algorithms to cluster, associate, and/or dimensionally reduce copious quantities of unlabeled data (IBM, 2019). Clustering pertains to the compartmentalizing of data into clusters which helps determine similarities and differences in the data. Unsupervised algorithms associate or derive the relationship between these clusters. Dimensional reduction is a tool used to reduces data input while maintaining data integrity. This allows for accurate and faster results when dealing with exceptionally large data sets.

Topic Modeling

Topic Modeling is a ML process designed to determine Topic Correlation. This type of unsupervised ML applies to evaluating textual data, correlation of topics and the significance of the correlation. Topic Modeling ML has been implemented through means of unsupervised, supervised and combinations of the two to aid in cyber security. Topic Modeling extracts topics from data and then determines the significance of the topics. There is a variety of options for algorithms to complete the topic extraction that will be discussed further in depth. Algorithms differ in how the topics are graded by metrics to decipher the correlation between them. Topic Modeling Algorithms are comprised of a process of data preparation and topic extraction. Topic Modeling Algorithms can include unsupervised, supervised or a method of combining unsupervised and supervised Modeling. The significant difference in the vectorization in Unsupervised Topic Modeling compared to supervised Topic Modeling is the algorithm identifies the extracted topics when topic categories are untrained (Krzeszewska et al., 2022). The distinction that supervised learning involves providing guided test samples that the algorithm can mimic and replicate (Shalev-Shwartz and Ben-David, 2014). This is contrast to unsupervised where there is no distinction between training and the test data provided for the algorithm. Before the categorization of either computational option, Text Preparation must be conducted to prepare the data to be assessed.

Topic Modeling has become an asset in the field of cyber security due to its ability to take in substantial amounts of data and sort it into clusters to classify the information. Due to its more widespread use, there has been an increasing trend in identifying how attackers formulate strategies and attack vulnerabilities in new and existing software. To identify where the attack is targeting, data is received from the ongoing attack and is sent to the Topic Modeling algorithm to identify matching terms and keywords of the attack. When the information is processed, it goes through multiple phases to generate information including data processing, text cleaning, bigram, and trigram models, building the topic model, visualizing the topics and associated keywords, and finally mapping the topics to OWASP Top 10 (Vanamala et al, 2019).

Topic Modeling research emphasized the effect of polysemy in Topic Modeling. The ability to properly categorize the true meaning of a word significantly impacts the results of Topic Modeling (Zhu et al., 2019). The researchers introduced the Joint Topic Word-Embedding (JWT) process. The model evaluated sentences to gather the true meaning of the words identified for classification. This allowed for proper locating of the word into proper topic categories. The authors achieve results significantly better than directly fine-tuning the pre-trained ELMo or BERT algorithms.

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the most common used algorithm for Topic Modeling. LDA is utilized for content analysis, designed for automated mapping, or organizing of large data archives of documents based on latent topics, measured as patterns of word co-occurrence. LDA is a latent variable model and a mixed membership model, where documents or data each belong to several different latent classes simultaneously (Vanamala et al., 2019). This prevents a one-dimensional classification of a document or set of data and allows more than one label to be applied to the data. LDA also considers the frequency of topics.

LDA-based Topic Modeling research conducted by Kim et al. (2022) evaluated 6755 journal articles. These journal articles spanned a 30-year range and was used to evaluate the change in topics associated with virtual reality in education. The researchers displayed percentages of topic frequencies to evaluate the change in percentage of these topics during four time periods during the 30-year total time span. Results from this LDA model included topic percentages of virtual reality-based education research could be largely divided into four periods: 1992–2011 (3.49%), 2012–2016 (18.96%), 2017–2019 (30.14%), and 2020–2022 (42.46%). Secondly, overall published research topics included education and educational research (16.33%), computer science (7.39%), surgery (4.13%), and psychology (4.03%). The third result of this research identified “Virtual Reality,” “Applied Science-Basel,” and “Interactive Learning Environments” were the most published journals for virtual reality-based education research. An identified limitation of this study was that the semantic analysis using LDA was insufficient and a metric to grade accuracy of the algorithm was not mentioned.

Sharma1 et. al. (2022) published research with findings of the implementation of LDA to determine research pattern trends in blockchain technology. The blockchain concept uses databases that record transactions over a network in a decentralized manner. IEEE, Springer, ACM, and other digital databases were used to accumulate 993 published papers for this research. The researchers used LDA to extract key terms and key documents found for each topic followed by an additional step to semantically map these topics. The most prevalent research topics identified by frequency in the study are “Taxonomy and Architecture of Blockchain,” followed by “Blockchain Implementation and Integration over various technologies.” The research limitation was addressed by adding the additional step to address semantics to get a clearer division of categories produced by LDA.

An LDA topic model was implemented to derive typical radar software defects to validate the effectiveness and applicability. A twenty four percent increase of accuracy for defect classification was achieved after the Bag of Words dictionary was altered with radar software requirements words and included additional weighting. This created a specialized word set for the LDA algorithm before topic identification. The accuracy and recall rate for this study were used to derive the increase in accuracy of an unaltered LSA model compared to the altered LSA model. The limitation of semantics was addressed with the alteration of Bag of Words, no other limitations were discussed by the researchers.

StreamFed-LDA, a distributed algorithm for LDA was proposed by Guo and Li (2021). The altered LDA model addressed the LDA deficiencies of learning evolving topics effectively, provides high-quality inference results and inferring the new data instances in real-time. NeurIPS, All-the-News, New York Times (NYTimes), and Public Medicine (PubMed) seven datasets were utilized in this experiment. The proposed model was compared to SVB-LDA model and resulted in a higher quality of topic consistency in repeated comparisons. This was achieved by implementing the ideal log-likelihood value, learned from SVI-LDA model. For inference quality, the researchers reduced the perplexity metric, which represents the randomness of each word that could be generated from vocabulary, by adjustments to the limited local update and step size. This resulted in a much lower perplexity rate compared to SVB-LDA. Running time of the training stage was used to evaluate latency comparison. The proposed model was able to provide more consistent lower runtime than the compared model.

Latent Semantic Analysis

Latent Semantic Analysis (LSA), another form of Topic Modeling, is a natural language processing algorithm designed to find the relationship between sets of documents and the terms they contain (Prakash et al., (2022). LSA is commonly used for extracting and representing the contextual meaning of words and to compute the similarity between words, sentences, or whole documents. LSA learns topics by making a matrix decomposition on a matrix of the document term using Singular Value Decomposition (SVD), a mathematical tool for analyzing matrices (Bellaouar et al., 2021).

Research focused on the retrieval of topics and their semantic relationship to enhance the results for the user of the search queries using LSA was conducted to improve document summarization (Al-Sabahi et al. 2018). Data sets used were DUC 2002, DUC 2004, and Multilingual 2015 Single-document Summarization. Word embeddings were utilized to enhance the LSA model input matrix weighting schemes. The results show that the proposed model improved the performance of the LSA algorithm in document summarization. Using metrics from the ROGUE automatic evaluation package for summarization, the proposed model outperformed two standard LSA models. The researchers disclosed that this model was not sufficient in producing shorter summaries of the documents tested.

Another research using LSA investigates the effectiveness of a teacher/student internet network by Ullah et al. (2020). LSA is used to compare semantics of a teachers question and the students answers or responses. The automation use of LSA provides evaluation of student responses by comparing the response to answer keywords that were predetermined for the question. The dataset used was from the LMS of Virtual University, Pakistan. The model was able to meet or exceed human accuracy and significantly improved the processing time. Similarity value was the metric used to compare the answer guide and student responses. This research was limited by incorporating a metric for evaluation of accuracy, the researchers intend to incorporate cosine similarity in the future.

A Heterogeneous-LSA model (h-LSA) was researched and developed to supersede standard LSA results for semantic space, dimensional reduction, and information retrieval (León-Paredes et al., 2017). The h-LSA model performed the indexing of text documents five thousand documents from PubMed Central (PMC) database with significantly reduced execution times with the use of multi-CPU and GPU architecture. The model was also evaluated for accuracy using Cosine Similarity. Execution rates achieved varied from 3.82 to 8.65 times faster than the standard LSA model used for comparison. The h-LSA model performed better in accuracy for two of three use cases compared to the standards LSA.

In 2020, An intelligent decision support system for software plagiarism detection was researched and implemented (Ullah et al., 2021). The researchers combined the use of term document matrix (TDM), the SVD algorithm, LSA and the synthetic minority over-sampling technique (SMOTE) method. After tokenization and frequency values were determined, TDM was used to determine TFIDF weighting to rank the tokens locally and globally. SVD was then utilized to decompose the three weighted matrices created in the previous step. LSA was then used to derive latent variables from the SVD records. The SMOTE method was then used to overcome the class imbalance when training features resulting in some improper categorization. This resulted in a 92.72% classification accuracy, outperforming five previous research results ranging from 2011 to 2019. The authors noted the limitation and future direction of reducing dependency of SVD use with LSA since SVD is complex to operate.

Sanguriet et al. (2020) proposed the addition of LSA to improve results of a previously researched method, Document Co-citation Analysis (DCA). The primary benefit of adding LSA allowed incorporation semantic similarity with a document similarity measure assigned to each document. Tourism supply chain metadata from the Scopus database was used for this research. With the addition of LSA, the researchers were able to increase the network and cluster analysis comparison of both the matrices. Limitations the researchers faced included a small sample size for data and time complexity.

Supervised Machine Learning

Unlike unsupervised ML which uses algorithms to analyze and cluster unlabeled data sets, supervised ML takes labeled datasets which are used to train algorithms into classifying data and predicting outcomes accurately. When looking at supervised ML it is often separated into two types of problems. The first is classification problems which are algorithms that are used to separate data into distinct categories. Some examples of algorithms that can do this include support vector machines (SVM), Decision Trees, and Random Forest. The other supervised ML type is regression problems which use algorithms to find relations between two types of variables. These types of models are used primarily to predict numerical values based on separate data points.

Support Vector Machine

SVM, otherwise known as support vector machines, is an algorithm used to find an optimal hyperplane within a N-dimensional space that can classify multiple points of data. To find this optimal hyperplane, we find a maximum margin or the maximum distance between data points of two classes and once found we can plot future points with more confidence outside of this maximum margin. Once the hyperplanes have been found, they are then used as decision boundaries to help classify data points being used. Depending on which side of the plane they fall on, they can be classified into that specific class. Along with the hyperplane there are also support vectors which help to position and orient the hyperplane and can help to maximize the margin which the points are classified at (Asim et al., 2021).

One use of the SVM algorithm has been with calculating and finding future information about

Covid-19 cases relating to deaths, new confirmed cases, and recovery cases. The overall goal of this research was to try to forecast the amount of people who can be infected by new cases and deaths including the expected recoveries for the upcoming 10 days. When calculating death rates, new infection cases, and recovery rates among individuals in 10 days, it was shown that compared to other methods used to calculate scores for these categories that the SVM algorithm performed the worst giving a R-squared score of 0.53 for deaths, 0.59 for new infections, and 0.24 for recovery rates. One problem that this research did run into however while using this method of machine learning was that SVM gave far worse accuracy's when given a smaller data set to train with and process. Based on the information from this research, although SVM was not the best algorithm used for it, it will be useful knowledge to know that it will be more accurate when given a larger data set to run according to Rustam et al. (2020).

Another case involves research which was conducted to find the performances of different machine learning techniques for disease prediction in patients by Uddin et al. (2019). Overall, the intended goal was to determine which articles used more than one machine learning method in order to calculate health predictions using two databases called Scopus and PubMed. Based on the findings, SVM was one of the most likely algorithms used when paired with more than one type of algorithm and SVM was the third most used algorithm with articles that used only one algorithm. Based on results from this research, SVM was shown to be the most frequently used algorithm.

A third research topic by Delli et al. (2018) found using the SVM algorithm is in determining the quality of the printing models. The overall goal of this research was to find methods which provides image processing and machine learning to monitor defects such as filament running out or defects to the prints structure. With the use of SVM, the researchers were able to create two categories of training models called good and bad which they used to train the algorithm. By checking RGB values within the print it was able to tell if the print was defective and if it was it would stop the printing process. Two limitations were found with this research the first was that the printing process had to be paused to take a clear image of the print along and the second was that defects weren't checked in the vertical plane of the print due to the limitations of just one camera. Based on the information in the research, it could be useful for us by providing an idea on how we could categorize training models.

A fourth research that uses the SVM algorithm includes a topic in detecting food in images is provided by McAllister et al. (2018). The primary goal of the research was to find the effectiveness of using deep feature extraction to classify food image datasets. With the use of SVM they were able to classify images into eleven categories and found a seventy eight percent accuracy when using it. Along with this classification however, they often paired it up with a secondary method in order to determine better outcomes with given data sets. Overall, their work in determining food based on images was a success with SVM finding a wide variety of accuracies with a max of around ninety eight percent. This could be useful to our research due to showing a wider variety of classification criteria compared to other algorithms that were used.

Finally, another research that used SVM in their work used it for population genetics research. The purpose of research by Schrider and Kern (2018) incorporated machine learning methods to track population genomic datasets due to their explosive increase in size. In this research, they were able to use SVM to determine and between the selective sweeps they wanted and neutrality. From the sweeps collected they were able to graph information on genomic position as well as relative values of statistics. Overall, they were able to use this method to tailor their needs in creating genomic predictors based on training sets that were passed.

Naïve Bayes

Naïve Bayes is a type of statistical classifier, which is used to identify probabilities of its target. With this type of method, it treats all data, or any variables as being mutually correlated and finds how likely it is to be related to a particular section of data. There is also a specific probability theorem that is used to calculate the probability of its type of class which is P(C|X) = (P(X|C).P(C))/P(X). To breakdown this theorem, P(C|X) is the posterior probability of the target class, P(C) is the prior probability of the class, P(X|C) is the likelihood of the predictor of the class, and P(X) is the prior probability of the predictor class.

One research topic that uses the Naïve Bayes algorithm involves its use in identifying disease prediction. Overall, their goal was to determine based on two databases called Scopus and PubMed which articles used more than one machine learning method to calculate health predictions. Based on their findings they found that Naive Bayes was the second most algorithm used when paired with more than one type of algorithm and was the second most used algorithm with articles that used only one algorithm. Based on results from this research, Naïve Bayes appeared to succeed well with large datasets and for taking less data for training. Some limitations that were found with it however were that classes were mutually exclusive and there was a presence of dependency between attributes in the classification (Uddin et al., 2019).

Research by McAllister et al. (2018) also utilizes Naïve Bayes algorithm is in detecting food items in images. The primary goal of the research was to find the effectiveness of using deep feature extraction to classify food image datasets. With the use of Naïve Bayes, they were able to find food within images at a rate of ninety eight percent which was the lowest among the types of algorithms used for the deep feature types. Overall, although Naïve Bayes gave the lowest percentage in their work, they were still able to determine food based on images. This could be useful to our research due to being able to use it with continuous values, so we are able to assume a normal distribution with the data.

A third research topic that uses this method includes research involving a study on liver disease (Rahman et al., 2019). The primary goal of this research was to predict results of liver diseases more efficiently and accurately in order to reduce the cost of diagnosis in the medical sector. When running tests with the machine learning algorithms they analyzed 583 liver patient's data to calculate their datasets. When running these tests with Naïve Bayes, the accuracy performs the worst at around 53%, the precision performs the worst at 36%, and Naïve Bayes also performed the worst when calculating logistics regression at about 53%. Although Naïve Bayes appeared to run the worst in the study, their best attribute was in calculating sensitivity in the dataset which was the highest among all the other machine learning techniques. Based on this information, it could be helpful toward our research by illustrating the strengths of what Naïve Bayes can provide when calculating information from a specific dataset.

Research by Ullah et al. (2021) involves software that searches for plagiarism in academia. The primary goal of this research was to find was to provide instructors with software to search for plagiarism between students' code due to the time-consuming process of checking it manually for grading purposes. In the document, they use Naïve Bayes as a method to check for plagiarism within a student work. They used four different training sets which were 50%, 60%, 70%, and 80% when they wanted to test the data. When they ran the process, they used the Naïve Bayes algorithm to detect similarities once the method was tokenized to break down the codes. Limitations found in this research primary were with issues when tokenizing the data and getting the breakdown of code ready for similarity detection. This information could be useful to our research due to understanding how we can use Naïve Bayes to test for similarity detection which we will be using to detect attacks on code from the MITRE database.

Random Forest

The Random Forest (RF) ML algorithm is an algorithm which consists of many individual tree data structures. Each tree in the RF has multiple classifications which when combined give the classification of the given data which was input into the RF. To determine the resulting classification, the forest chooses the individual trees that contain the data or classification that is the most like each data point and outputs that as the resulting conclusion for the set. Once you have your chosen data set collected for the tree, each tree then randomizes the number of attributes which represent the nodes and leaves of a regular tree. After this, each tree is then grown to its maximum depth and width without any points being removed (Livingston, 2005).

RF key benefits include reduced risk of overfitting, provides flexibility, and ease of determining feature importance (Bellaouar et al., 2021). Compared to Decision Trees method, RF creates many trees which reduces the overall variance and prediction error. RF provides flexibility by being able to manage regression and classification tasks. RF provides an increased ability to evaluate variable importance. Mean Decrease in Impurity (MDI) and mean decrease accuracy (MCA) are variables RF evaluate well. RF key challenges are data resources, time consumption of the process and interpretation complexity. With increased number of trees (data structures), a larger resource for data is required. More data structures also increase the processing time. Deriving final results from multiple structures does add complexity to determining final results.

One research topic that uses random trees involves its use in identifying disease prediction. Overall, their goal was to determine based on two databases called Scopus and PubMed which articles used more than one machine learning method to calculate health predictions. Based on their findings, they found that RF was the second least used algorithm used when paired with more than one type of algorithm and was the fifth most used algorithm with articles that used only one algorithm. Based on results from this research, RF was shown to be less likely used within health research which used machine learning. Based on this information, it was found that RF provides less chance of variance in the training data and happens to scale well for larger data sets. Some limitations that came with it though include a higher expense in computation to run and the base classifiers need to be defined beforehand (Uddin et al., 2019).

Another research topic that uses random trees involves research in identifying food items in images (McAllister et al., 2018). The primary goal of the research was to find the effectiveness of using deep feature extraction to classify food image datasets. With the use of RF, they were able to find food within images at a rate of 99% when used with the deep feature types. Overall, RF did an excellent job in identifying types of food although in some data sets it did struggle to differentiate some types of foods like small grains. This information could be useful to our research due to being able to use it to lower the variance within the results calculated.

Research by Schrider and Kern, 2018 also test this method for the population genetics research. RF was used to distinguish between recombination rate classes and illustrate motifs or sequences within the collected data. Although this method worked as intended, they were able to find a faster method with greater computational efficiency, which they decided to switch to. From this information, this research can help us by illustrating the use of RF to find unique sequences among massive amounts of data which needs to be classified.

A fourth research topic of this method includes research involving a study on liver disease (Rahman et al., 2019). The primary goal of this research was to predict results of liver diseases more efficiently and accurately to reduce the cost of diagnosis in the medical sector. When running tests with the machine learning algorithms they analyzed 583 liver patient's data to calculate their datasets. When running RF in this study, it was found that the accuracy of the data was around 74%, the precision was 85%, and the sensitivity was 81%. Overall, RF provided stable accuracy when calculating the proper data of true positives, false positives, false negatives, and false positives.

A final research topic of this method includes research involving a study of multiple sclerosis with MRI's conducted by Sweeney et al., 2022. The primary goal of this research was to determine how classification algorithms, feature extraction functions, and the interplay between both classification algorithms and feature extraction functions affected the performance of lesion segmentation methods. Based on the study, they found that RF performed better than other simpler algorithms when calculating the rates of true positive test rates to false positive test rates. When interpreting the data, the algorithm provides little intuition about underlying classification problems and only provides the computational complex rules for making predictions (Sweeney et al., 2014).

Neural Networks and Deep Learning

Neural Networks are sets of mathematical models that are used to simulate the thinking of a human brain. These models have three separate processes, which include multiplication, summation, and activation. First, the inputs are multiplied by a certain weight. Next in the process, the sum of all the inputs that have been weighted which is then compared to a certain bias. Finally, after comparing the weighted sum to the bias, an activation sequence is used to compute the final product. When each input is weighed, it represents its synapse strength. When the synapse strength is high that means there is a higher weight and when the strength is low that means that it has a lower weight. Along with this, weights can also be positive or negative values. Negative weight means that there is inhibited neuron activity while a positive weight strengthens that activity (Mohamed, 2017).

One research topic with uses of neural networking involves its use in identifying disease prediction. Overall, their goal was to determine based on two databases called Scopus and PubMed which articles used more than one machine learning method to calculate health predictions. Based on their findings, they found that neural networks were the third most used algorithm used when paired with more than one type of algorithm and was the most used algorithm with articles that used only one algorithm. Based on results from this research, neural networks were shown to be highly involved within health research. Based on this information, it was found that neural networks benefited due to its advantage in complex nonlinear relations and application to classification and regression problems used by the data sets. Although they do have these benefits, some limitations found when using it were that it has characteristics of being like a black box and is computationally extensive when you train it (Uddin et al., 2019).

Another research topic that that uses neural networking was in identifying food items within images. The primary goal of the research was to find the effectiveness of using deep feature extraction to classify food image datasets. By using an adaptive learning rate as well as one thousand iterations they were able to identify food with an accuracy of around 99.4% Based on the information from this research it could be useful for us to later implement due to its ability to iterate through massive datasets (McAllister et al., 2018).

A third research idea that shows this method participates in a study on population genetics (Schrider and Kern, 2018). The purpose of this research was to use machine learning methods to track population genomic datasets due to their explosive increase in size. In the research, neural networks were used to learn the mapping of summary statistics onto parameters in an efficient manner due to the savings in computational cost. Overall, this method was the most widely used one due to its outputs in categorical and continuous parameters. When comparing what we are doing for our research to this research topic, it is beneficial to see the wide range of data neural networking will be able to benefit when looking at the wide variety of population genetics.

A fourth research topic of this method being used includes research involving a study on multiple sclerosis with MRI's. The primary goal of this research was to determine how classification algorithms, feature extraction functions, and the interplay between both classification algorithms and feature extraction functions impacted the performance of lesion segmentation methods. Based on the study, neural networks were found to perform better than other algorithms due to their decision boundaries. When taking data, the performance is stable at around 3000 voxels however, once it reaches a higher amount of around 15000 voxels the data on the unnormalized graphs scaling becomes unstable and jumped to different scaling values inconsistently. One limitation found using this method in the research however was its behavior of classifiers on unbalanced data due to its poor performance. Based on this research, we can learn from the limitations found when given an unbalanced dataset and find ways to decrease performance errors when passing new data into the algorithm (Sweeney et al., 2022).

IV. Discussions

Semantics of words have a crucial role in properly categorizing words through ML. Two different words can be processed into the same word, which potentially provides inaccurate classification of several words. An example of preprocessing of the word desert and deserted, these words both become desert. The meaning of the word deserted becomes lost. To properly identify and recommend the most relevant vulnerabilities from CAPEC, a ML process efficient in semantics will produce the best outcome. With recommending cyber security vulnerabilities, semantics is of utmost importance to incorporate in the ML process. The next discussion is the consideration of implementing an unsupervised, supervised or a semi-supervised ML model. This research is aims to compares key words from one singular document (SRS document) and compares them to a key word list generated for each CAPEC vulnerability. The CAPEC data base is quite large and incredibly detailed and specific.

Unsupervised ML algorithms are primarily used for separation of data into cluster, exploring relationships between data and dimensionality reduction. Dimensionality reduction may be a useful tool to reduce the Large CAPEC data set while aiming to maintain integrity of the data. LDA research revealed a large amount of adaptation needed to achieve desirable results to overcome semantical flaws. The LSA algorithm is designed to incorporate semantics and identify the relationships between vectors that words are separated into. Through extensive research, LSA is commonly combined with SVD or another algorithm that is complex in nature. Unsupervised methods in general do not have metrics available to accurately assess the accuracy of the model making interpretation of the effectiveness of results more complex.

Supervised ML is a less complex process and requires less tools than unsupervised ML (IBM, 2019). Supervised ML uses a training dataset, which helps derive accurate results more timely compared to unsupervised ML, which needs a large unlabeled data set to learn. Classification of data in organized into specific categories through a trained dataset compared to unsupervised ML, which clusters objects into like groups identified by the algorithm. The largest limitation for supervised ML requires obtaining the training data set to prep the implemented algorithm. Supervised ML also is significantly more proficient at obtaining metrics for accuracy of results.

V. CONCLUSION

After identifying the importance of cyber security vulnerability controls during the requirement phase, the CAPEC software vulnerability repository was identified as the most useable repository for this research. The organization of the attack patterns allows for proper identification and allows for referral back to CAPEC for recommended defense strategy. Recent research in similar cyber security projects is reviewed and discussed. Topic Modeling, unsupervised and supervised ML methods are defined, and examples of recent research and the applicability of these methods are evaluated. Through the development of this research, our future work will include implementing supervised machine learning. The CAPEC repository offers a prelabeled data set which will assist in implementation of a trained data set. Supervised ML also has the advantage of higher proficiency in use of metrics to refine the ML process, allowing for evaluation and process refinement to enhanced results. A training set for the SRS document will need to be created or located to perform supervised ML. Without a known related research comparison utilizing supervised ML, our future work will test and compare results from Naïve Bayes and RF ML methods. Naïve Bayes statistically performs well with large and small data sets, which will suit the smaller data set of the SRS documents and the larger data set for the CAPEC Vulnerabilities. The RF capability to prevent over fitting is desirable for the complex data from CAPEC. The algorithm that produces the greatest number of accurate recommendations of CAPEC attack patterns from an SRS document will be used to implement an automated tool to process ad visualize the results.

VI.REFERENCES

[1] Vanamala, M., Y. Xiaohong, and B. Kanishka. 2019. Analyzing CVE Database UsingUnsupervised Topic Modelling. 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Dec 05-07, IEEE Xplore Press, USA, pp: 72-77. DOI:10.1109/CSCI49370.2019.00019.

[2] Vanamala, M., J. Gilmore, X. Yuan, and K. Roy. 2020. Recommending Attack Patterns for Software Requirements Document. 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020, IEEE Xplore Press, USA, pp: 1813-1818. DOI:10.1109/CSCI51800.2020.00334.

[3] Vanamala, M., X. Yuan and K. Roy. 2020. Topic Modeling And Classification Of Common Vulnerabilities And Exposures Database. 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Aug 06-07, IEEE Xplore Press, South Africa, pp: 1-5. DOI:10.1109/icABCD49160.2020.9183814.

[4] Kanakogi, K., H. Washizaki, Y. Fukazawa, S. Ogata, T. Okubo, T. Kato, H. Kanuka, H. Hazeyama and N. Yoshioka. 2022. Comparative Evaluation of NLP-Based Approaches for Linking CAPEC Attack Patterns from CVE Vulnerability Information. Applied Sciences, 12 (7): 3400. DOI:10.3390/app12073400.

[5] Krzeszewska, U., A. Poniszewska-Marańda and J. Ochelska-Mierzejewska. 2022. Systematic Comparison of Vectorization Methods in Classification Context. Applied Sciences 12 (10): 5119. DOI:10.3390/app12105119.

[6] Alyami, H., M. Nadeem, A. Alharbi, W. Alosaimi, M. Ansari, D. Pandey, R. Kumar and R. Khan. 2021. The Evaluation of Software Security through Quantum Computing Techniques: A Durability Perspective. Applied Sciences, 11 (24): 11784. DOI:10.3390/app112411784.

[7] Guru Prasad, G., M. Badrinarayanan and C. Ceronmani Sharmila. 2022. Efficacy and Security Effectiveness: Key Parameters in Evaluation of Network Security. International

Journal of Performability Engineering, 18 (4) : 282. DOI:10.23940/ijpe.22.04.p6.282288.

[8] Zhu, L., Y. He, and D. Zhou. 2020. A Neural Generative Model for Joint Learning Topics and Topic-Specific Word Embeddings. Transactions of the Association for Computational Linguistics, 8: 471–485. DOI:10.1162/tacl_a_00326

[9] Asim, M., M. Ghani, M. Ibrahim, W. Mahmood, A. Dengel, and S. Ahmed. 2021. Benchmarking Performance of Machine and Deep Learning-Based Methodologies for Urdu Text Document Classification. Neural Computing & Applications, 33 (11): 5437. DOI:10.1007/s00521-020-05321-8.

[10] Bedi, G. 2018. Simple Guide to Text Classification(NLP) Using SVM and Naive Bayes with Python. Medium. https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34 (Accessed on November 17, 2022)

[11] Shalev-Shwartz, S., and S. Ben-David. 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. ISBN: 1107057132. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

[12] Macsai, D. 2012. The most important company you've never heard of. 1 Minute Read. Fast Company. https://www.fastcompany.com/3017927/30mitre (Accessed on November 10, 2022)

[13] A course module on HTML5 new features and security concerns

[14] Vanamala, M., Yuan, X., & Morgan, M. (2019). A course module on HTML5 new features and security concerns. Journal of Computing Sciences in Colleges, 34(5), 23-30.

[15] Forest-[Frederick-Livingston].pdf (Accessed on November 12, 2022)

[16] Vanamala, M., Yuan, X., Smith, W., & Bennett, J. (2022). Interactive Visualization Dashboard for Common Attack Pattern Enumeration Classification. ICSEA 2022, 79.

[17] Mohamed, A. 2017. Comparative study of four supervised machine learning techniques for classification. International Journal of Applied Science and Technology, 7 (2): 1-15. https://www.ijastnet.com/journal/index/859

[18] Uddin, S., A. Khan, M. Hossain, and M. Moni. 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19 (1): 1-16. DOI:10.1186/s12911-019-1004-8.

[19] Delli, U., and S. Chang. 2018. Automated process monitoring in 3D printing using supervised machine learning. Procedia Manufacturing, 26: 865-870. DOI:10.1016/j.promfg.2018.07.111.

[20] McAllister, P., H. Zheng, R. Bond, and A. Moorhead. 2018. Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. Computers in Biology and Medicine, 95 : 217-233. DOI:10.1016/j.compbiomed.2018.02.008.

[21] Schrider, D., and A. Kern. 2018. Supervised machine learning for population genetics: A new paradigm. Trends in Genetics, 34(4): 301–312. DOI:10.1016/j.tig.2017.12.005

[22] Rahman, A., F. Sazzadur, F. Shamrat, Z. Tasnim, J. Roy, and S. Hossain. 2019. A comparative study on liver disease prediction using supervised machine learning algorithms. International Journal of Scientific & Technology Research, 8 (11): 419-422. http://www.ijstr.org/final-print/nov2019/A-Comparative-Study-On-Liver-Disease-Prediction-Using-Supervised-Machine-Learning-Algorithms.pdf

[23] Lasky N, Hallis B, Vanamala M, Dave R and Seliya N, (2022,November) Machine Learning Based Approach to Recommend MITRE ATT&CK Framework for Software Requirements and Design Specifications.In The 4th Colloquium on Analytics, Data Science, and Computing (CADSCOM 2022).ACM.Prakash, A., N. Singh, and S. Saha. 2022. Automatic extraction of similar poetry for study of literary texts: An experiment on Hindi poetry. ETRI Journal, 44 (3): 413-425. DOI:10.4218/etrij.2019-0396.

[24] Bellaouar, S., M. Bellaouar, and I. Ghada. 2021. Topic modeling: Comparison of LSA and LDA on scientific publications. In 2021 4th International Conference on Data Storage and Data Engineering, February, pp. 59-64. DOI:10.1145/3456146.3456156.

[25] Al-Sabahi, K., Z. Zuping, and Y. Kang. 2018. Latent semantic analysis approach for document summarization based on word embeddings. KSII Transactions on Internet and Information Systems, 13 (1): 254-276. DOI:10.3837/tiis.2019.01.015.

[26] Ullah, F., J. Wang, M. Farhan, S. Jabbar, M. Naseer, and M. Asif. 2020. LSA based smart assessment methodology for SDN infrastructure in IoT environment. International Journal of Parallel Programming, 48 (2): 162-177. DOI:10.1007/s10766-018-0570-1.

[27] Kim, D., and T. Im. 2022. A Systematic Review of Virtual Reality-Based Education Research Using Latent Dirichlet Allocation: Focus on Topic Modeling Technique. Mobile Information Systems, Volume 2022. DOI:10.1155/2022/1201852.

[28] Sharma, C., and S. Sharma, S. 2022. Latent DIRICHLET allocation (LDA) based information modelling on BLOCKCHAIN technology: a review of trends and research patterns used in integration. Multimedia Tools and Applications, 81:36805-36831. DOI:10.1007/s11042-022-13500-z.

[29] Guo, Y., and Li, J. 2021. Distributed Latent Dirichlet Allocation on Streams. ACM Transactions on Knowledge Discovery from Data (TKDD), 16 (1) : 1-20. DOI:10.1145/3451528.

[30] León-Paredes, G., Barbosa-Santillán, L., and Sánchez-Escobar, J. 2017. A heterogeneous system based on latent semantic analysis using GPU and multi-CPU. Scientific Programming Techniques and Algorithms for Data-Intensive Engineering Environments, Volume 2017. DOI:10.1155/2017/8131390.

[31] Ullah, F., Jabbar, S., and Mostarda, L. 2021. An intelligent decision support system for software plagiarism detection in academia. International Journal of Intelligent Systems, 36 (6): 2730-2752. DOI:10.1002/int.22399.

[32] Sanguri, Kamal, Atanu Bhuyan, and Sabyasachi Patra. 2020. A semantic similarity adjusted document co-citation analysis: a case of tourism supply chain. Scientometrics, 125 (1): 233-269. DOI:10.1007/s11192-020-03608-0.

[33] CAPEC, 2022. Common Attack Pattern Enumeration and Classification (CAPECTM). https://capec.mitre.org (Accessed on August 23, 2022)

[34] MITRE ATT&CK®, 2022. https://attack.mitre.org Accessed 8/23/2022.

[35] CVE, 2022. https://cve.mitre.org (Accessed on August 25, 2022)

[36] CISA, 2019. What Is Cybersecurity? | CISA. https://www.cisa.gov/uscert/ncas/tips/ST04-001. (Accessed on September 14, 2022)

[37] NIST, 2019. About NIST. https://www.nist.gov/about-nist. (Accessed on September 21, 2022)

[38] IBM, 2019. What is machine learning?

https://www.ibm.com/topics/machine-learning?lnk=fle. (Accessed on September 2022)

Cite this article as :

M Vineela, Thota Prathyusha, Ette Pravalika, "Emotion based Music Recommendation System", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.294-299, March-April-2023.