Proceedings of the Second International Conference on Artificial Intelligence and Smart Energy (ICAIS-2022)
IEEE Xplore Part Number: CFP22OAB-ART; ISBN: 978-1-6654-0052-7

Email Spam Classification Using Machine Learning Algorithms

Dr.P.Vishnu Raja
Dept of CSE
Kongu Engineering College
Perundurai
pvishnu@kongu.ac.in

Dr.K.Sangeetha
Dept of CSE
Kongu Engineering College
Perundurai
sangeetha_k@kongu.ac.in

G.SuganthaKumar
Dept of CSE
Kongu Engineering College
Perundurai
suganth200131@gmail.com

R.Varun Madesh
Dept of CSE
Kongu Engineering College
Perundurai
varunmadesh8@gmail.com

N.K.K.Vimal Prakash
Dept of CSE
Kongu Engineering College
Perundurai
vimalprakash.621@gmail.com

2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) | 978-1-6654-0052-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAIS53314.2022.9743033

***Abstract*— Email is the easiest way to communicate worldwide today. To get e-mail junk mail, the preceding set of rules compares every e-mail message with junk mail records earlier than generating receivers. In the project, an electronic mail acquisition machine primarily based totally on SVM development has been advocated and more NB is used in the proposed system. With word spreads, the length of the word meaning, the proportion of word stops can also be seen. To highlight the electronic mail junk mail trend, a singular version that enhances the arbitrary production of the detector of SVM and NB algorithm using both spam and non-spam spaces. Theater analysis and test results show that the performance detection of the enhanced SVM and NB is higher.**

***Keywords*— Spam, Ham, classification, Support Vector Machine, Naïve Bayes ,K-Nearest Neighbor**

I. INTRODUCTION

Detecting and fixing broken equipment is a critical part of resolving any network security threats. Because it can be used to execute a variety of assaults, including denial of service attacks, spam, stealing user identities, and spreading malicious software, among others. Spam is one of the most serious risks, with attackers launching single attacks on as many devices as possible in the network. Although a few technologies, such as spam signatures and spam behaviour analysis, have helped to solve problems in the past, they no longer operate on huge networks. Furthermore, these approaches do not allow for the transmission of spam over the internet. This paper also discusses existing programs and associated difficulties. For monitoring and identifying spam attacks on a real-time network, the tool's design and implementation are critical. A mechanism is being created in this project to distinguish between spam and non-spam devices by exchanging internet messages. The application maintains track of everything. Machine and keeps track of the network's spam proportion. From the network administrator's perspective, this protects the privacy of customers that exchange non-spam emails with their code.

The most common spam that most users encounter on a daily basis is email spam. Anonymity, bulk email, and unsolicited emails are three aspects of email spam. Anonymity is a property of a different encryption that conceals the email's sender. Unsolicited emails are transmitted to uninvited receivers, and emailing is described as sending repeated identical emails to a huge number of groups. Email spam is defined as an email sent to several groups without any request for anonymity. This is the most prevalent type of spam that many users see on different blogs. Spammers utilize blog postings to drive spam victims to spam websites. In search engines, the number of such blogs is gradually increasing. It's mostly used to promote search services like Wikipedia, blogs, and guest books, among other things.

II. RELATED WORKS

In this work of B.Biggio, G.Fumera, I.Pillai, and F.Roli et.al [1], " A survey and experimental evaluation of image spam filtering techniques," has proposed junk that is depending on your location. All of the recommended techniques have the same goal: to keep image spam out of our inboxes. Spam senders use a variety of strategies to circumvent screening, and each requires analysis to establish where the filters should be placed in order to defeat the tricks and prevent spam senders from filling our mailbox. Different strategies are created through different techniques.

In the work of A.Heydari, M.A.Tavakoli, N.Salim, and Z. Heydari, et.al [2], "Detection of review spam: A survey" suggested that the proliferation of E-commerce sites has made the web an excellent source for collecting customer reviews about products; as there is no level control anyone can write anything that leads to spam reviews. This project previews and reviews extensive research on how to detect spam. In addition it provides an artistic environment that reflects some previous effort to learn spam detection. Today due to the popularity of Ecommerce sites it has become a target for spammers without known email and web spam. Update spam refers to spam

written spam so that they can denigrate product features or defile themselves.

In the work of T.Oda, T.White et.al. [3], " Increasing the accuracy of a spam-detecting artificial immune system," suggested Spam, which is equivalent to electricity and junk mail, affects about 600 million people globally. Senders adjust to guarantee that their messages are noticed, whilst anti-junk mail answers evolve to lessen the quantity of junk mail dispatched to users. The use of an antibody model to effectively defend email users from spam is examined in this research.

In this work of A.H.Mohammad, R.A.Zitar et.al [4], "Application of genetic optimized artificial immune system and neural networks in spam detection" Has proposed The present thesis dreams to make an in-depth study of adaptive identification, digital channel equalization, beneficial link artificial neural community and Artificial Immune Systems . These new models are performed for adaptive identification of complex nonlinear dynamic plant life and equalization of nonlinear digital channel. Investigation has been made for identification of complex Hammerstein models. The results of simulation are compared with those acquired with FLANN-GA, FLANN-PSO and MLP-BP based totally completely hybrid approaches. Improved identification and equalization basic overall performance of the proposed method have been placed in all cases..

In this work of A.Visconti, H.Tahayori et.al [5], "Artificial immune system based on interval type-2 fuzzy set paradigm" has proposed Spam has made the mail gadget unreliable due to the fact mail may be stuck falsely with the aid of using unsolicited mail filtering earlier than being introduced to the recipient conversely unsolicited mail mails may be discovered withinside the recipient mail box. Artificial Immune System version is stimulated from the herbal immune gadget. A getting to know segment is supplied to praise the cells that efficaciously understand the unsolicited mail e-mails. Instead of the usage of clonal choice, bad choice is used withinside the schooling segment which ended in better checking out overall performance because of the decreased quantity of detectors. The herbal immune gadget defends the frame towards dangerous illnesses and infections. It is able to spotting and removing any overseas molecular or molecule.

In this work of J.Balthrop, S.Forrest, M.R. Glickman et.al [6], "Revisiting LISYS: parameters and normal behavior" Has proposed this challenge research a simplified shape of LISYS, an synthetic immune gadget for community intrusion detection. Different experiments via way of means of extraordinary organizations have executed markedly extraordinary effects in this problem, possibly due to variations many of the models, algorithms, and records sets. This variety of formalisms and experimental test-beds are viewed as positive, withinside the experience that at this early degree it's miles crucial to discover a own circle of relatives of immune-stimulated algorithms in place of standardizing too early on arbitrary choices.

In this work of S.Forrest, A.S.Perelson et.al [7], "Self-Non-self-Discrimination in Computer" Has proposed The purpose of this paintings is to examine emergent mechanisms that would probably be embedded in multi-sellers structures which will resolve complicated troubles of disbursed prognosis. An software become applied to pupil prognosis withinside the context of an academic surroundings for studying simple programming skills. One fundamental purpose of this paintings has been to examine emergent mechanisms that would probably be embedded in multiagent structures which will resolve complicated troubles of disbursed prognosis. In this area we've applied some operational multiagent prognosis structures alternating sellers' behaviors from reactive to cognitive approaches .

In the work of M.Gong, J.Zhang, J. Ma, and L. Jiao et.al [8], "An efficient negative selection algorithm with further training for anomaly detection" has proposed the negative selection algorithm is one of the oldest immune inspired classification algorithms and was originally intended for anomaly detection tasks in computer security. After initial enthusiasm, performance problems with the algorithm lead many researchers to conclude that negative selection is not a competitive anomaly detection technique. However, in recent years, theoretical work has led to substantially more efficient negative selection algorithms. Here, the results of the first evaluation of negative selection with r-chunk and r-contiguous detectors that employs these novel algorithms have been reported. On a collection of 14 datasets from real-world sources, negative selection is compared with r-chunk and r-contiguous detectors against techniques based on kernels, finite state automata, and n-gram frequencies, and find that negative selection performs competitively, yielding a slightly better average performance than all other techniques investigated. Because this study represents, to our knowledge, the most comprehensive one of string-based negative selection to date, the widely held view that negative selection is not a competitive anomaly detection technique may be inaccurate.

In the work of A.Bratko, B.Filipič, G.V.Cormack, T.R.Lynam, and Zupan et.al [9],"Spam filtering using statistical data compression models" Has proposed Spam filtering poses a unique trouble in textual content categorization, of which the defining function is that filters face an energetic adversary, which continuously tries to stay away from filtering. Since unsolicited mail evolves constantly and maximum sensible packages are primarily based totally on on-line consumer feedback, the venture requires rapid, incremental and sturdy mastering algorithms. By modeling messages as sequences, tokenization and different error-susceptible preprocessing steps are overlooked altogether, ensuing in a way this is very sturdy. The fashions also are rapid to assemble and incrementally updateable. Electronic mail is arguably the «killer app» of the internet. It is used each day with the aid of using tens of thousands and thousands of human beings to talk around the world and is a mission-essential utility for lots businesses. Over the remaining decade, unsolicited bulk e-mail has end up a chief trouble for e-mail users.

In the work of J.Gordillo, E.Conde et.al [10], "An HMM for detecting spam mail" has proposed Spam filtering poses a

unique trouble in textual content categorization, of which the defining feature is that filters face an energetic adversary, which continuously tries to prevent filtering. By modeling messages as sequences, tokenization and different error-inclined preprocessing steps are disregarded altogether, ensuing in a way this is very robust. The fashions also are rapid to assemble and incrementally updateable. Electronic mail is arguably the «killer app» of the internet. It is used every day with the aid of using hundreds of thousands of human beings to talk around the world and is a mission-essential utility for lots businesses. Over the final decade, unsolicited bulk e-mail has grow to be a first-rate trouble for e-mail users.

In the work of I.Idris, and A.Selamat et.al [11], "Improved email spam detection model with negative selection algorithm and particle swarm optimization" Has proposed Spam filtering poses a unique hassle in textual content categorization, of which the defining feature is that filters face an lively adversary, which continuously tries to stay away from filtering. By modeling messages as sequences, tokenization and different error-inclined preprocessing steps are unnoticed altogether, ensuing in a technique this is very robust. The fashions also are rapid to assemble and incrementally updateable. Electronic mail is arguably the «killer app» of the internet. It is used each day via way of means of hundreds of thousands of human beings to talk around the world and is a mission-vital utility for lots of work.

III. SYSTEM ANALYSIS

A. EXISTING SYSTEM:

In existing system email spam classification has used all machine learning classification algorithms for email spam detection they can get the accuracy between 85% - 95% described in Fig 1. only, but with Bio inspired and Particle swarm optimization algorithms they used to improve the accuracy of the models to increase accuracy from 95% to 98%.

Classifiers	Average
IBK	85.79%
OneR	81.91%
Naïve Bayes	90.46%
Naïve Bayes Multinomial	92.65%
SMO	93.98%
AdaBoost	89.48%
Bagging	89.37%
ZeroR	63.07%
Decision Stump	81.33%
Hoeffding Tree	84.33%
J48	89.53%
Random Forest	93.04%
Random Tree	83.13%
Naïve Bayes Multinomial Text	63.07%

Fig 1. Existing models accuracy without optimization

B. PROPOSED SYSTEM:

SVM and multinomial NB are used in the proposed system. With the distribution of words, mean word length, stop word ratio can also be identified. The proposed system Email spam is a convenient means of communication throughout the entire world today. The increased popularity of spam emails in texts and images requires a real-time protection mechanism for multimedia flow. In this project "Email Spam classification" is implemented using the CRISP-DM methodology which increase the classification accuracy without using optimization techniques.

C. MACHINE LEARNING MODELS

Naive Bayes and SVM. Email spam classification done using traditional machine learning techniques comprise Naive Bayes and SVM (support vector machines), due to less time to train. Also, not opting for neural algorithms due to less data and computing resources. Both are good at handling large number of features; in the case of text classification each word is a feature and there are thousands of words based on the vocabulary of the corpus. SVM works best with high dimensional data, a vocabulary with 1000 words means each text in the corpus will be represented with a vector of 1000 dimension.

When there are sufficient number of features, both SVM and Naïve Bayes can work with less data as well.

Naïve Bayes does not suffer from curse-of-dimensionality because it treats all features as independent of one another. Also, one of the benefits of features being independent is: For example, most spam emails contain words such as money and investment, etc, but it is not necessary that all the mails containing both words money and investment are considered to be spam.

IV. PROPOSED WORK

This research will experiment SVM and NB Machine learning models.

The paper aims to achieve:

1) To explore machine learning algorithms for the spam detection problem.

2) To investigate the workings of the algorithms with the datasets.

3) To implement and increase accuracy of machine learning algorithms without optimization techniques.

4) To test and compare the accuracy of the models with existing papers.

5) To implement the framework using Python.

Scikit-Learn library will be explored to perform the experiments with Python[12], and this will conduct pre-processing, training and testing to calculate the results. The program scripts will be implemented without the optimization techniques and compared with the previous results.

V. APPROACH TO THE STUDY

The CRISP-DM methodology is used to implement "Email Spam categorization" in this project. The steps of Business Understanding, Data Understanding (Data Description and Exploration), Data Preparation, Modelling, and Evaluation will be covered. The project is built utilising a Python class object-based approach. Naive Bayes and SVM (Support vector

machines) system learning methods are used to detect email spam. It also covers the whole programme flow for implementing a Python-based email spam classifier, including Data Retrieval, Data Visualization, Data Preparation, Modelling, and Evaluation. Most of us think of spam emails as unwanted messages that are repeatedly sent for the aim of advertising and brand promotion. Such email addresses canbe banned continuously, but it is ineffective because spam emails continue to be common. Fraudulent e-mails, identity theft, hacking, viruses, and malware are just a few of the major types of spam emails that pose a significant security concern. In order to deal with spam emails, a robust real-time email spam classifier has to be built, that can efficiently and correctly flag the incoming mail spam, if it is a spam message or looks like a spam message. The latter will further help to build an Anti-Spam Filter. Google and other email services are providing utility for flagging email spam but are still in the infancy stage and need regular feedback from the end-user. Also, popular email services such as Gmail, Yandex, yahoo mail, etc provide basic services as free to the end-user and that of course comes with EULA. There is a great scope in building email spam classifiers, as the private companies run their own email servers and want them to be more secure because of the confidential data, in such cases email spam classifier solutions can be provided to such companies.

In this project “Email Spam classification” is implemented using the CRISP-DM methodology. You will get to know Business understanding, Data Understanding (Data Description and Exploration), Data Preparation, Modelling, and Evaluation steps. Project is implemented using Python class object-based style. Email spam detection is done using Naive bayes and SVM (Support vector machines) machine learning algorithms. Further, it shows the complete program flow for Python-based email spam classifier implementation such as Data Retrieval Flow, Data Visualization Flow, Data Preparation Flow, Modelling, and Evaluation Flow. Most of us consider spam emails as one which is annoying and repetitively used for purpose of advertisement and brand promotion. Such email-ids can be blocked but it is of no use as spam emails are still prevalent. Some major categories of spam emails that are causing great risk to security, such as fraudulent e-mails, identify theft, hacking, viruses, and malware. In order to deal with spam emails, a robust real-time email spam classifier has to be built, that can efficiently and correctly flag the incoming mail spam, if it is a spam message or looks like a spam message. The latter will further help to build an Anti-Spam Filter. Google and other email services are providing utility for flagging email spam but are still in the infancy stage and need regular feedback from the end-user. Also, popular email services such as Gmail, Yandex, yahoo mail, etc provide basic services as free to the end-user and that of course comes with EULA. There is a great scope in building email spam classifiers, as the private companies run their own email servers and want them to be more secure because of the confidential data, in such cases email spam classifier solutions can be provided to such companies. “alternately” (unless you really mean something that alternates).Fig 1. System flow diagram describes about the work flow of the project.

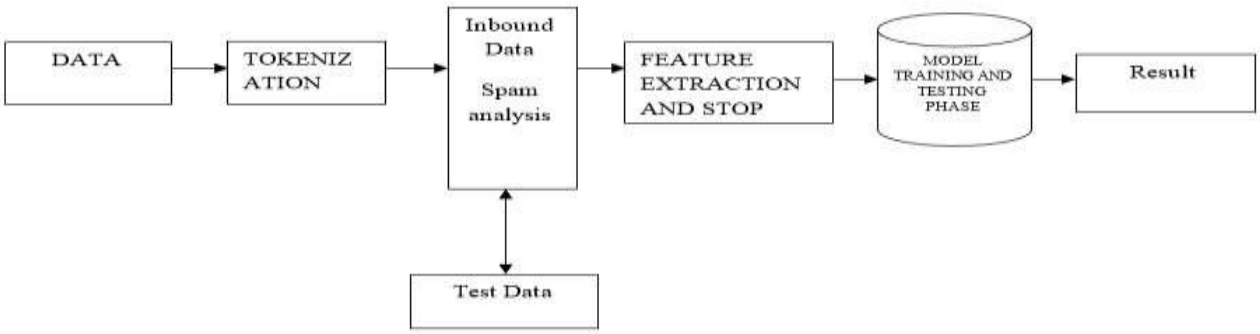

Fig. 2. System Flow Diagram

A. Dataset:

The dataset are accessed from public repository Kaggle for this project and included each email as an individual text file. The text files were string based. The name of the dataset is spam filter which contain 5000 – 6000 unique values in text format, with 0 or 1 as shown in Fig 3. diagram. In the pre-processing the spam and ham text files based on the values are split. Splitting the data into training and test datasets, where

training data as 80 percent and test as 20 percent.

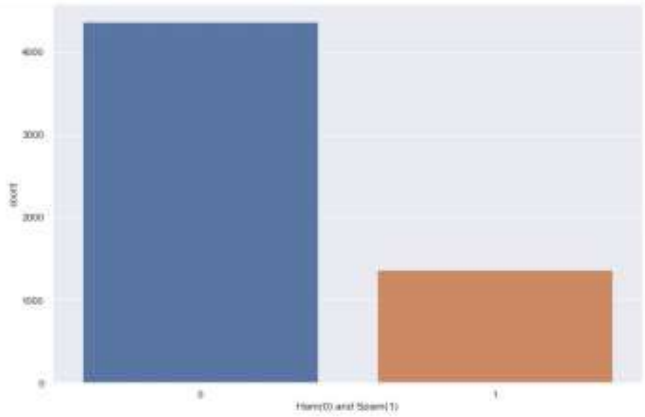

Fig 3. Ham(0) and Spam(1) Count

B. *Tokenization:*

Tokenization is the process of breaking down sentences in an email Into character words (tokens). These tokens are kept in a list and used in test data to determine where each word in the email came from. This will aid Algorithms in figuring out whether or not an e-mail need to be labelled as junk mail or ham.

C. *Feature Extraction and Stop Words:*

This become used to delete undesired phrases and characters from every email, in addition to build a word bag that could be compared between algorithms. While counting, the Scikit-learn 'Count Vectorizer' module adds This become used to delete undesired phrases and characters from every email, in addition to words, such as A, In, The, Are, As, Is, and others, because they are ineffective in determining if an email is spam or not as in Fig 4. After that, this sample is prepared for the vocabulary learning programme. the size of the Count vector matrix is 5728 x 20114, where 5728 represents the number of documents in the corpus and 20114 represents the number of features in the vocabulary.

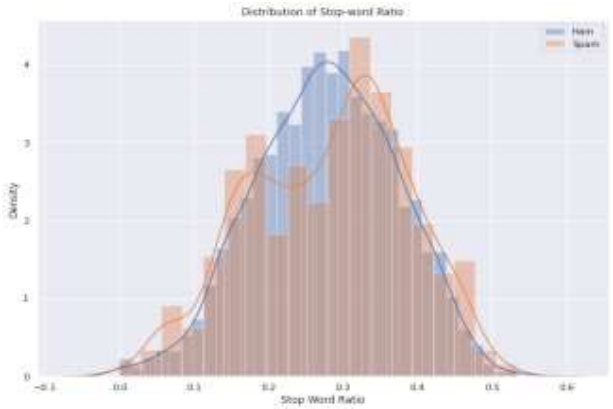

Fig. 4. Distribution of Stop-word ratio

D. *Model Training and Testing Phase:*

Supervised learning and Naïve Bayes methods were employed, and the version became educated with regarded statistics and examined with unknown statistics to expecting accuracy and different performance indicators, as stated in the research. The accuracy in results were obtained using the CRISP-DM method in machine learning models and word cloud. The word cloud is a data visualization technique used for representing text data as shown in Fig 5.

Fig 5. Spam Word Cloud Image

VI. RESULT

To take a look at and evaluate the usefulness and accuracy of SVM and NB with modern NS approaches, many measures can be used. After that, in data mining, a mathematical quality measurement can be applied to machine learning and journals. Sensitivity (SN), Positive Predictability (PPV), Poor Predictability (NPV), F-measure (F1), Accuracy (ACC).

A. Sensitivity (SN):

Positive pattern proportions that are accurately recognized by positive are measured by sensitivity.

$$SN = \frac{TP}{(TP + FN)}$$

where TP denotes the number of true positives and FN denotes the number of false negatives.

B. Positive prediction value (PPV):

The test is result positive predictive value tells you how likely you are to pass the test. average true percentage of total total number of patterns recognized as positive. Measuring opportunities in a well-predicted pattern such as positive,

$$PPV = \frac{TP}{TP + FP}$$

where the number of false positive is FP.

C. Negative prediction value (NPV):

The test's negative predictive value also provides the proportion of genuine negatives for the entire number of negative patterns.

$$NPV = \frac{TN}{FN + TN}$$

Where the number of true negative values is TN.

D. Accuracy (ACC):

The percentage of correctly categorised samples is the accuracy[12] measure as shown in Fig 6.

$$ACC = \frac{TP + TN}{TN + TP + FP + FN}$$

E. F-measure (F1):

Both sensitivity and positive predictive value are combined in the F-measure[13] as shown in Fig 6.

$$F1 = 2 \times \frac{PPV \times SN}{PPV + SN}$$

(a)

	Naïve Bayes
Accuracy Score	0.9947
F1 Score	0.9886

(b)

	SVM
Accuracy Score	0.9842
F1 Score	0.9685

Fig. 6. (a) Naive Bayes Accuracy and F1-Score, (b) SVM Accuracy and F1-Score

F. Experimental setup:

Here,Table 2 show the effectiveness of the improved SVM AND NB method.Spam messages and ham messages (not spam) are identified using 99% high-speed svm-nb and knn accuracy u*p to 95%.*Table 1 describes the efficiency of SVM & NB compared to other algorithms of existing paper algorithms.

Table 1. Efficiency table

Algorithm	Efficiency
SVM & NB	99%
SGD	97%
MNB	98 %
DT	94%
RF	93%
KNN	95%

VII. REFERENCES

[1] B. Biggio, G. Fumera, I. Pillai, and F. Roli, "A survey and experimental evaluation of image spam filtering techniques", Pattern Recognition Letters,

vol 32, no. 10, pp.1436-1446, 2011.

[2] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey", Expert Syst. Appl., vol. 42, no. 7, pp. 3634-3642, 2015.

[3] T. Oda, T. White, "Increasing the accuracy of a spam-detecting artificial immune system", in: The 2003 Congress on Evolutionary Computation (CEC), 2003.

[4] A.H. Mohammad, R.A. Zitar, "Application of genetic optimized artificial immune system and neural networks in spam detection", Appl. Soft Comput., vol. 11, no. 4, pp. 3827–3845, 2011.

[5] A. Visconti, H. Tahayori, "Artificial immune system based on interval type-2 fuzzy set paradigm", Appl. Soft Comput., vol. 11, no. 6, pp. 4055–4063, 2011.

[6] J. Balthrop, S. Forrest, M.R. Glickman, "Revisiting LISYS: parameters and normal behavior", in: Proceedings of the Congress on Evolutionary Computing, 2002.

[7] S. Forrest, A. S. Perelson, " Self Nonself Discrimination in Computer ", 1994.

[8] M. Gong, J. Zhang, J. Ma, and L. Jiao, "An efficient negative selection algorithm with further training for anomaly detection", Knowl.-Based Syst., vol. 30, pp. 185–191, 2012 .

[9] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and Zupan "Spam filtering using statistical data compression models", J. Mach. Learn. vol7, 2673–2698,

[10]J. Gordillo, E. Conde, "An HMM for detecting spam mail", Expert Syst. Appl., vol. 33, no. 3, pp. 667–682, 2007.

[11]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, ''Scikit-learn: Machine learning in Python,'' J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

[12]Bashar, Abul. "Survey on evolving deep learning neural network architectures." Journal of Artificial Intelligence 1, no. 02 (2019): 73-82.

[13]Vijayakumar, T. "Comparative study of capsule neural network in various applications." Journal of Artificial Intelligence 1, no. 01 (2019): 19-27.

[14]Chen, Joy Iong-Zong, and Kong-Long Lai. "Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert." Journal of Artificial Intelligence 3, no. 02 (2021): 101-112.

[15]Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3, no. 2 (2021): 95-113.

[16] Manoharan, J. Samuel. "Study of Variants of Extreme Learning Machine (ELM) Brands and its Performance Measure on Classification Algorithm." Journal of Soft Computing.