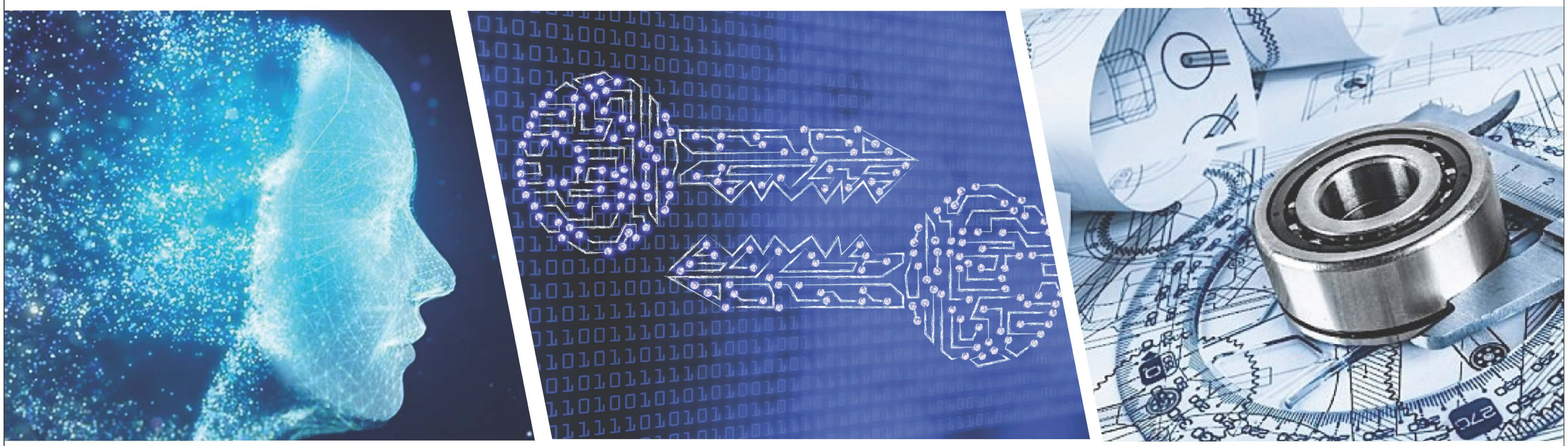
Volume 3, Issue 1

ISSN: 2706-7351

Foundation University Journal of Engineering and Applied Sciences

Volume 3 Issue 1



Foundation University Journal of
Engineering and Applied Sciences
E-mail:- editor.fujeas@fui.edu.pk

FUJEAS
Biannual

Volume 3, Issue 1

ISSN: 2706-7351

Foundation University Journal of Engineering and Applied Sciences

Foundation University Journal of
Engineering and Applied Sciences
E-mail:- editor.fujeas@fui.edu.pk

FUJEAS
Biannual

Aims of the Journal

The Foundation University Journal of Engineering and Applied Sciences (FUJEAS) is a biannual peer-reviewed research journal published by Foundation University Islamabad, Pakistan. The journal aims to be a leading cornerstone of researchers , academicians and practitioners who are interested in high quality new knowledge, product development, issues and challenges in the field of Engineering, Applied Sciences and Computer Science. FUJEAS welcomes research in relation to (but not limited to) the following areas:

- Ad Hoc Networks for Security
- Context-Aware Computing
- Advance Computing Architectures
- Bioinformatics
- Broadband and Intelligent Networks
- Broadband Wireless Technologies
- Cloud Computing and Applications
- Communication Systems
- Cryptography
- Computational Intelligence
- Embedded Systems
- Information System in Health Care
- Information Processing
- Information Systems and Applications
- Internet Technologies
- Big Data
- Data Analysis
- Data Mining
- Data Retrieval
- Digital Signal Processing Theory
- Emerging Signal Processing Areas
- Social Media Analysis
- Evolutionary Computing
- Fuzzy Algorithms
- Internet of Things
- Information Retrieval
- Human Computer Interaction
- Image Analysis and Processing
- Multidimensional Signal Processing
- Multimedia Applications
- Neural Network
- Information and Data Security
- Information Management
- Internet Applications and performances

Review Policy of the Journal

The Foundation University Journal of Engineering and Applied Sciences highly promote the vision of Higher Education Commission (HEC) of Pakistan and implements anti-plagiarism policy. Plagiarism in all its form constitutes unethical publishing behavior and is not acceptable. The journal carries out a plagiarism tests and reserves the right to remove and retract a plagiarized article.

Submission and acceptance of papers depend upon the response of the reviewers. Plagiarism checking, and review process may take from four weeks to eight weeks depending upon the response from the reviewers and compliance from authors. After satisfactory reviews, the editorial board accepts the paper from publication and acceptance letter is issued to the author(s).

Editorial Board

TABLE OF CONTENTS

Foundation University Journal of Engineering and Applied Sciences
Vol. 3, Issue 1, 2022.
DOI: 10.33897/fujeas.v3i1.564

Research Article

Urdu Sentiment Analysis Using Deep Attention-Based Technique

Naeem Ahmed[1], Rashid Amin[1,2*], Huma Ayub[3], Muhammad Munwar Iqbal[1], Muhammad Saeed[1], Mudassar Hussain[4]

[1]Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan
[2]Department of Computer Science, University of Chakwal, Chakwal, Pakistan
[3]Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan
[4]Department of Computer Science, University of Wah, Wah Cantt, Pakistan
*Corresponding Author: Rashid Amin. Email: rashid.amin@uoc.edu.pk

Abstract:

Sentiment analysis (SA) is a process that aims to classify text into positive, negative or neutral categories. It has recently gained the research community's attention because of the abundance of opinion data on the internet. Deep learning techniques are widely used for language processing but are seen as black boxes, and their effectiveness comes in interpretability. The major goal of this article is to create an Urdu SA model that can comprehend review semantics without the need of language resources. We design an attention-based neural network for the review level Urdu SA. For better results, we used a transfer learning approach that uses pre-trained embedding's. The Visualization of attention weights is also measured that uncovers the black box of the models and confirms their intuition, which aids in the interpretation of the model's learned representations. The proposed model is tested and evaluated in terms of accuracy and F1 score. The proposed model archives 91% accuracy and 88% F1 score, respectively.

Keywords: Deep learning; Urdu language; Natural language processing; Sentiment analysis

1. Introduction

Sentiment analysis (SA) tries to summarize attitudes about a particular subject, service, or entity by manipulating textual data of people or group of people. Sentiment classification categories text inputs into one of two categories (positive or negative) Yang, L, et al., (2020).Where SA is useful is the trend analysis, social net analysis, public opinion catagorization, policy evaluation, and decision-making direction. Currently, various machine learning (ML) models i.e., support vector machine (SVM), Random forest Naïve bayes (NB), Decision tree etc. are being used for sentiment analysis of various languages. learn hierarchical data Deep learning techniques are also being used for SA of different languages with huge amounts of data Soleymani, M, et al., (2017). Neural network models can representations by incorporating several processing layers. Emotion detection (ED) and analysis Asghar, M, et al., (2017) is used to determine the writer's implicit emotions from text. Various tasks like speech, facial expressions, posture, gestures, and text have all been used to identify emotion. Different approaches for emotion identification have been studied, such as the keyword methodology, which focuses on keywords in the text and includes a parser and an emotion lexicon. In a hybrid technique, classification models are trained utilizing extensive linguistic knowledge from lexicons, thesauri, and keyword and learning approaches Almani, N.M., et al., (2020).

Sentiment detection is important in affective computing Cambria, E, et al., (2017) (AC), which is focused with developing computer systems that can recognize and respond to human affective states Khotima, D.A.K, et al., (2018). Affect sensitive systems are used in a variety of fields, including education, gaming, psychological health, and customer service.

Urdu is a widely spoken language globally, with over 100 million speakers. Many people have recently started using Urdu in their tweets, reviews, and comments. As a result, sentiment analysis for the Urdu language has grown important Daud, A., et al., (2017). Several approaches and methods are available for text mining and sentiment analysis in English. However, sentiment analysis in other languages such as Urdu, Hindi, Arabic, and others has rarely been studied. The Urdu language's complex morphological structure and distinct script from English are two primary reasons for its lack of tools Mukhtar, N, et al., (2020).

Natural language processing (NLP) neural networks are often implemented for various tasks to improve accuracy. Each sentence is considered a collection of tokens, i.e., characters or words Khalid, K, et al., (2017). The ability of neural networks to extract multi-level representations of input text allows them to control local and long-range dependencies. In various field like voice recognition, computer vision, NLP, and other fields, neural networks abstract information from raw input. One-dimensional Convolutional Neural Networks are used in these networks to model time delayed sequential data.

Transfer learning is also used for solving various problems and allows applications for NLP, bioinformatics and computer vision Tan, C, et al., (2018). The main aim of transfer learning (TL) models is to build effective, reliable and accurate models in the target domain by using the data of source domin. Because of this technique we can minimize the training time and enhamce the accuracy of model easily. This technique is often used when we want high-performing models but have less data or when training data is expensive.

SA models are facing various challenges. These challenges include sarcasm detection, negations, compound phrases, repetition of words etc. Like some other languages, the Urdu language is widely used by individuals for data sharing on internet. It is clear from the literature study that techniques used in SA for other languages cannot be used to deal with Urdu language. Urdu sentiment analysis is becoming popular since few years because of its increasing rate on internet. The main objective of this study is to analyze and investigate the attention based deep learning technique for Urdu sentiment classification.

This study proposes a deep attention approach for generating Urdu review representations that do not rely on external resources like handmade features or sentiment lexicon. The study's first phase data is scraped from various blogging and social media websites. These websites include daily Pakistan, hamariweb.com, BBC Urdu, and many other Urdu blogs. Some part of the standard IMDB is translated into Urdu using google trans library. This data is preprocessed to tokenize and normalize the Urdu sentences. The reviews length was fixed to 500 words for effective model development. In order to generate Urdu review representations without the need of outside resources like hand-crafted features or sentiment lexicon. The study suggests a deep attention technique by using a transfer learning approach to generate successful models. In addition, a review-level sentiment analysis algorithm is designed to look at the polarity of reviews by choosing the most informative phrases that support a particular sentiment in any review. The models' "black box" is also revealed via visualizing attention weights, which supports users' intuition and facilitates the understanding of the model's learnt representations.The key contributions of the study is listed below:

- For sentiment analysis of Urdu, we suggest an attention-based deep learning model.
- Creating large Urdu data sets using various blogs data for Urdu SA.
- The use a deep attention approach for generating Urdu review representations that does not rely on external resources.
- To use the transfer learning technique for building effevtive Urdu SA model by selecting the most informative phrases that reflect a certain sentiment in a review.

Visualizing attention weights for understanding the most informative terms for uncovering the BlackBox of DL models. This study uses the transfer learning approach to develop effective models. To begin, the acquired data set is preprocessed to normalize and tokenize it. Secondly, a review-level sentiment analysis algorithm is designed to examine the polarity of reviews by selecting the most informative phrases that contribute to a certain sentiment in any review. The experiments were carried out using a transfer learning method that included pre-trained embeddings. Visualizing attention weights also uncovers the black box of the models and confirms their intuition, which aids in the interpretation of the model's learned representations. The paper is further divided into following sections. Section 2 discusses the related work for the proposed problem. The problem statement and formulation are discussed in section 3 while section 4 discusses the insights about the data set used in the study. Section 5 presents the proposed methodology and the results are explained in section 6, while section 7 concludes the article.

2. Related Work

Formerly, machine-learning approaches were used to apply several algorithms to the issue of classification tasks of sentiments. Bibi, R, et al., (2019). proposed a supervised machine learning technique for Urdu SA. They used a decision tree algorithm for the Urdu SA. Their approach is composed of two main phases. In the first phase, data were preprocessed, and unnecessary data was removed. Stop words, hashtags, and other unnecessary words were removed in this step. In the next phase, feature vectors were generated. For this purpose, positive and negative comments and POS tags were identified. Finally, the decision tree algorithm is applied to the data for sentiment classification. The Urdu tweets were used in this study as a data set. The proposed decision tree model got 90% accuracy.

Mukhtar, N, et al., (2018) presents research that focuses on sentiment analysis on the Urdu language. They gathered data from various blogs and preprocess it. After preprocessing, supervised machine learning techniques, i.e., Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), were used for sentiment classification. After comparing the performance of these models, they cannot get satisfactory results. Then feature extraction was analyzed, and 152 features were extracted. Then, after classifiers' training, they got a 67% accuracy level by using the KNN algorithm. So, in this study KNN outperform SVM and decision tree in sentiment classification of the Urdu language.

Syed, A.Z, et al., (2014)present a sentiment analysis article using sentiUnits. This article uses the identification of SentiUnits from the data by using shallow parsing. SentiUnits are expressions that have information about the sentiment expressed in the sentence. Lexicon of Urdu language is made in this article and then used for sentiment classification. This paper highlights the linguistic, i.e., grammar and morphology of the Urdu language and technical aspects of this problem. The evaluation of the system is done with various test text data and got satisfactory results. This article can be used as a baseline for sentiment analysis.

Mukund, S, et al., (2012) present research on sentiment by using Urdu blog data. This research used structural correspondence learning to transfer Urdu sentiment data from Urdu news wire to Urdu blog data. As newswire data is written in Latin script and has code-mixing and code-switching behaviour, the data from these two platforms are not trivial. So, for making pivots, two oracles were used in this study. First was the Transliteration oracle, and the one was called translation oracle. The transliteration oracle was for script variation, while the translation was used for code-switching and codemixing behavior. They introduce a new part-of-speech tagging process that helps them identify words based on POS categories, representing code-mixing behavior. They evaluate their model against a supervised learning model and compare results based on various performance measures. After evaluation, they got a 59.4% precision value and 62.4% recall value for the proposed technique.

Mehmood et al. proposed an Urdu sentiment analysis system using RUSA data set Mehmood, K, et al., (2019). The data set contained 11,000 reviews of products. They presented three distinct techniques to achieve text normalization. RUSA dataset is used to establish the BL accuracies in the technique. The second and third research utilized six phonetic algorithms and TERUN to optimize the RUSA data set. The resultent data was used for the training of ML models. According to the empirical review, the

results of TERUN were statistically significant and comparable to those obtained by phonetic algorithms. The TERUN word normalization technique was then generalized from a corpus-specific to a corpus-independent technique. The study concludes that text normalization enhances machine learning algorithms' accuracy rate. Another result was that a phonetic algorithm designed for one language will not generalize well to other languages unless it becomes properly updated to fulfill the phonological needs of its target languages.

Nasim, Z, et al., (2020) present an Urdu SA article combining various linguistic and lexical features. Their work focuses on building of Urdu SA system for Urdu tweets. A Markov chain model was used to design the approach in this paper. The help of Twitter API gathered the data set. The proposed model was trained on that data and model was able to predict people's attitudes based on their tweets. They also discussed about the challenges and limitations of Urdu SA systems. Their proposed model accurately predicted positive emotions because of less positive tweets in the data set.

Naqvi, U, et al., (2021) discussed many potential approaches available for sentiment analysis, but little work has been done on analyzing Urdu sentiments. This paper discusses the increasing rate of Urdu language on internet and need of Urdu SA systems. Their article outline and summarises the most recent SA updates and classification techniques used in the Urdu language. Various improvements were suggested in this article for Urdu SA. Figure 1 shows some proposed techniques for sentiment analysis of various languages. Masood et al. Masood, M, et al., (2022) used deep learning techniques to classify Urdu sentiments using a custom data set. The data set used in this study is 3995 Urdu sentences having 3 sentiment catagories. They used LSTM model with 830 stem Urdu words after the preprocessing phase. After preprocessing the padding is performed for equlizing the length of vectors for training deep learning model. The results of this study shows that the proposed LSTM model achived 86% and 89% accuracy and F1-score repectively.

A word level translation framework was proposed by Asghar, M, et al., (2019) to enhance the Urdu SA

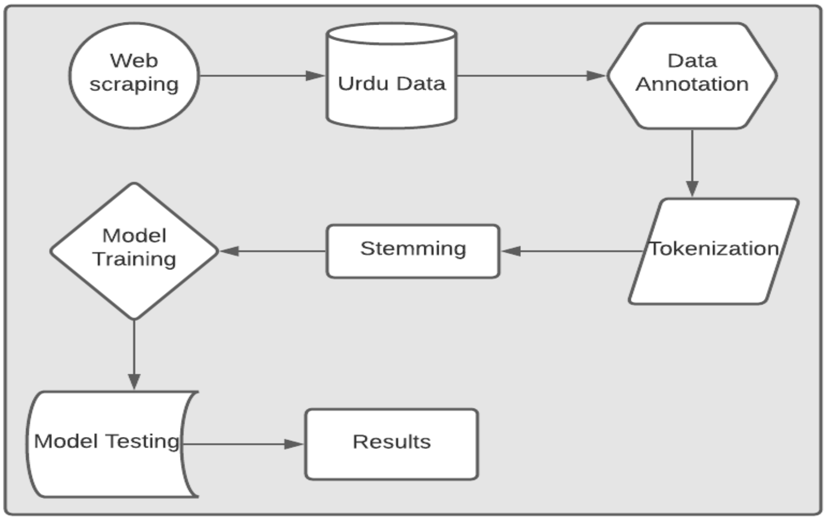


Figure 1:Techniques used for sentiment analysis

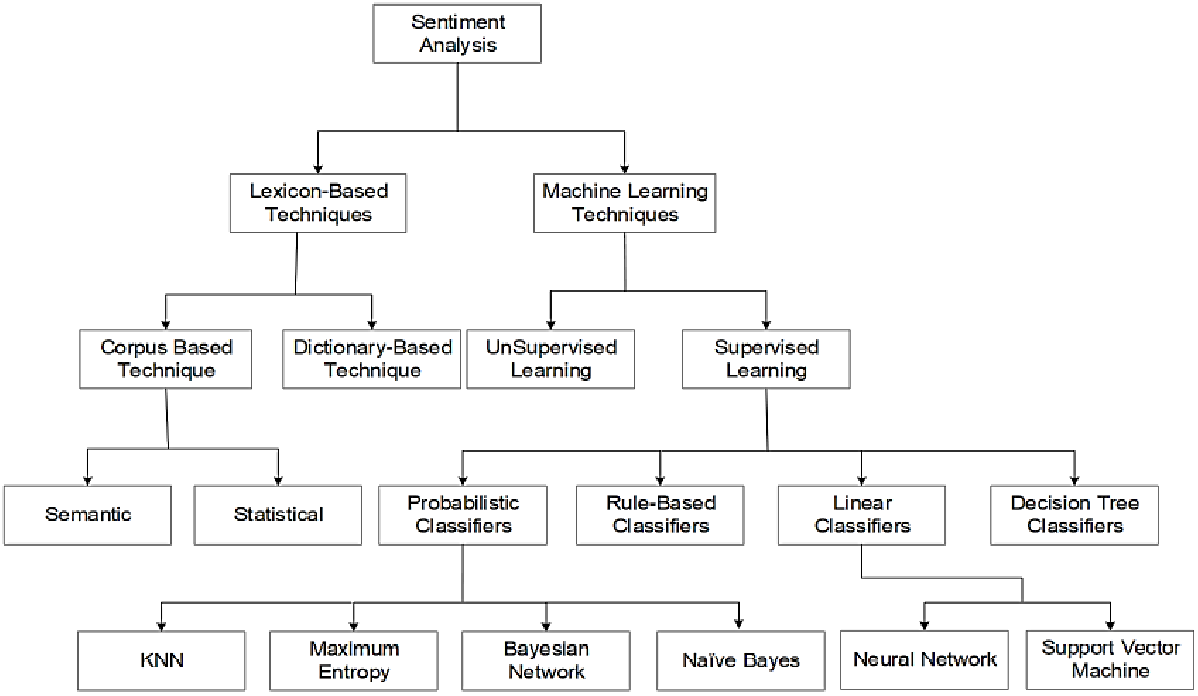


Figure 2: Process of sentiment analysis

lexicon. The framework was developed by combining different linguistic and lexicon resources, such as the English word list, SentiWord Net, the bilingual English-to-Urdu dictionary, Urdu grammar improvements, and a novel scoring mechanism. Their model consisted on three major modules, i.e., the collection of Words in English for an opinion, the translation of English words into Urdu, and sentiment scoring using SentiWordNet and manual scoring.

Hashim, F, et al., (2016) proposed a sentence-level Urdu SA method using nouns. They used Urdu news data for their lexicon-based method for Urdu SA. Urdu nouns are used for the detection of sentiments in sentences. Their proposed technique got 86.8 % accuracy and testing data set.

Naqvi, U, et al., (2021) Proposed deep learning techniques with different word vectors for Urdu sentiment analysis. They used various DL models, i.e., LSTM, CNN, BiLSTM and attention-based BiLSTM, to classify Urdu test sentiment analysis. In the sequential models, they used stacked layers and for CNN they used various filters with a single convolution layer. This study also analyzed the role of pretrained embeddings on an emotion classification problem. The attention-based Bi-LSTM model outperforms other models by achiving 77% accuracy and 72% F1-score in this study.

3. Problem Statement and Formulation

The proposed process of finding the polarity of piece of text can be explained as: Suppose, a review R_i=(W1+W2+W3,...Wn), which is made up of a words "w" that are padded to a range of "n" using padding tokens. Without employing any constructed features, the suggested method uses a binary classification system to determine the catagory of any review.

The suggested problem is a deep attention-based neural model that can identify significant and insignificant terms in any review. The model takes some distributed vector representations as input for terms, model will give the entire review's distributed vector representation, which will be divided into positive or negative classes using a linear classifier.

The initial stage in any NLP system should be text preparation and normalization Liu, H, et al., (2018). This phase removes noise from the data, including repeated and non-Urdu text such as URLs. The normalization procedure also transforms similar words into the same form. After that, all words are tokenized by the help of tokenizer. The step result is an integer sequence representing the input terms, with the complete data set encoded as integer values. In addition, all reviews are fixed to the same size by using the padding technique for shorter reviews.

4. Dataset

There are not as many standard data sets for the Urdu language. A very few data sets are available for the concerned language, but they are very small. So the data is gathered from various blogging and social media websites. These websites include daily Pakistan, hamariweb.com, BBC Urdu, and many other Urdu blogs. Some part of the standard IMDB Kumar, H, et al., (2019) reviews data set is also translated into Urdu for building a larger data set for the training and evaluation of the proposed models. The data collection from these websites is done using the "Beautiful soup" Patel, J, et al., (2020) python library and stored in a .csv file. The final data set contains more than 25000 rows and two columns. The classification of the data set is shown in Figure 3.

Twenty percent of the data set is used for the testing phase, 10% is used for validation, while 70% is used for training the proposed models. The size of every comment is set to 500 words, as this was determined to be the average length of comments in data set. To decrease the length reviews shorter than the mean value were padded, and reviews bigger were shortened.

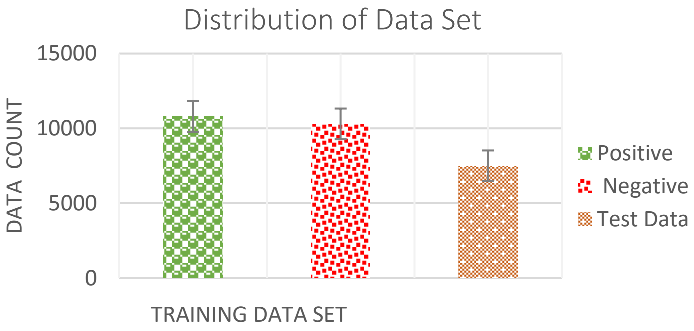


Figure 3: Distribution of dataset

5. Proposed Scheme

The model's primary objective is to investigate how attention mechanisms are employed to identify the most meaningful terms contributing to the polarity of overall Urdu reviews. The proposed approach attempts to address the limits of global sentence representation by focusing on and emphasizing smaller data blocks, in this case review terms.The proposed process for Urdu sentiment analysis is illustrated in Figure 4.

The proposed model is consisting on various layers. In start, the model allows distributed representation by employing an embedding layer. This layer passes input distributed representation of words to a GRU-based layer Zhang, Z, et al., (2018) which creates reviews' hidden representations. Warping the entire sentence into one single vector is not feasible. We are trying to represent excessive information in a restricted space and RNN layers cannot maintain the dependencies for more than a few time steps. As a result, to build a short-term memory, the attention layer is placed on top of the GRU layer, which retains the most important details from the sequence. The proposed of model flow is illustrated in Figure 4.

The attention layer adds the GRU hidden representations for every word depending on the specified weights for the final vector representation. The proposed model's attention layer build a distributed vector representation by using the essential terms of the data item. The weight is calculated at every hidden state at each time stamp is represented by "h_{it}". "T" is the number of time steps in the input, while "wit" is the hidden representation. The attention layer's weights, "i_w" and "b_w," are tuned during training to assign higher weights to a phrase's most essential words.

This phenomenon is illustrated in (1):

$$W_{it} = Tanh(I_w * h_{it} + b_w) \qquad (1)$$

The updated hidden representation for the current word "Wit" is then passed to the Softmax method, which returns the normalized important weight α_{it}.

$$\alpha_{it} = \frac{\exp{(W_{it})}}{\Sigma_{i=1}^{j} \exp{(W_{jt})}} \qquad (2)$$

In the end, this representation is constructed as a weighted sum of the word annotations depending on the weights. It may be considered a high-level illustration of the informative words used in any review.

$$F_{it} = \Sigma_{i=1}^{j}(\alpha_{it} * W_{it}) \qquad (3)$$

For the classification of reviews, the resultant vector representation "F" is transferred to a fully connected layer having a sigmoid activation function.

5.1 Model Details and Experiments

The transfer-learning technique is used at the embedding layer level in this study. The proposed model's embedding layer is weighted using a pre-trained model's final layer Naqvi, U, et al., (2021). As a result, the model inherits several common word patterns from the pre-trained model, and all it requires to discover additional relationships among words to categorize reviews. Various experiments were designed and assessed to see how effective pre-trained models were at solving the problem of Urdu SA. The baseline and the proposed model are trained and evaluated for the word embedding initialization with various test cases.

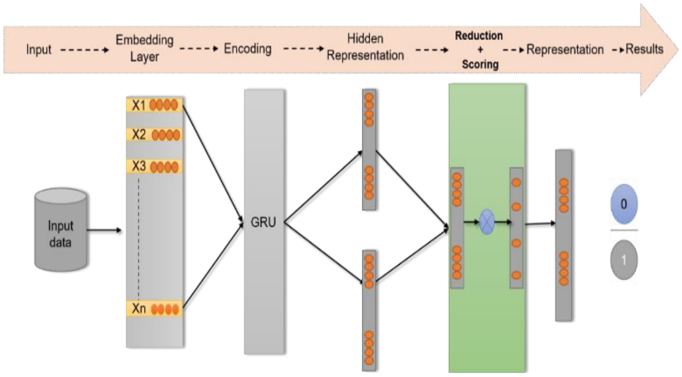


Figure 4: Proposed model for Urdu SA

The GRU layer and the final review representation on the GRU unit’s dropout are set to 0.25, whereas between layers, it is set to 0.35. Table 1 shows the values of parameters for the proposed model.

Table 1: Proposed model parameters

Parameter	Values
Epochs	30
Batch Size	32
Loss Function	Cross_entropy
Optimization Function	Adam
Learning Rate	0.001
Output layer function	Sigmoid
Train test ratio	70%-30%

The first scenario is to use the weights of the retained model to apply transfer learning to the current model. No pre-trained models were employed in the second scenario. Glorot and Bengio Glorot, X, et al., (2011) uniform initialization were used to set the weights of the embedding layer. The experiments are carried out on a Dell PC with a 3.4GHz processor and 16GB of RAM. The suggested model is implemented using Python 3.8.8 by the help of the Keras package.

6. Results and Discussion

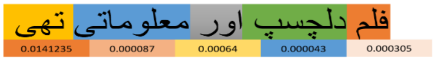
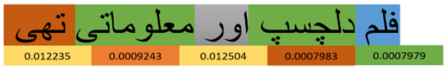
The proposed model output vector and attention weights vectors are presented for examing the black box of deep learning model's output. This was done to find the saliency characteristics contributing to the final classification. This is a missing feature in all sentiment classifiers for Urdu language. The approach utilized to assess reviews is divided into two categories. First, test the model's capability to identify the polarity of the entire review. The model's potential to determine the most informative terms in the given text is examined in the second stage.

Word context test was done on the test data to see whether the model could find the right class for the same term in different contexts or not. As shown in Table 2, for the reviews tested during this phase, the word “معلوماتی” might have a positive or negative sentiment.

After analyzing the proposed model, we identified that " فلم" is the most silent word in the review presented in Table 2. However, without considering the scores vectors that are given by the scoring function of the attention layer beacuse we cannot ensure that the final representations is chosen by the most meaningful terms that determine the outcome of model. As shown in Table 2 one of the least informative terms in the review in the SAL scoring vector is " معلوماتی". The output vector of SAL for the proposed model also have the same important rank, but if we analyze the SAL score presented in Table 2, it can be seen that words that come before and after " اور" have almost the same scoring range. The existence of " اور " leads to the conclusion that the model may build a language's linguistic structure by predicting that the two terms have the same meaning.

The both models were trained for 30 epochs and tested for the evaluation purposes. The attention based proposed model got the accuracy of 91% while the baseline model archives 86% accuracy. The models are

Table 2: SAL score

Original Label (Positive)	**Predicted Label (Positive 0.78)**
	The movie was intresting and informative.
SAL Output	 فلم دلچسپ اور معلوماتی تھی 0.0141235 0.000087 0.00064 0.000043 0.000305
SAL Score	 فلم دلچسپ اور معلوماتی تھی 0.012235 0.0009243 0.012504 0.0007983 0.0007979

also tested with F1-score values in which the proposed attention based model archives 88% F1-Score with pertained embedding while baseline model archives 85% F1-Score. The results of proposed attention based model is presented in Figure 5. As illustrated in Figure 5, the proposed model performance is low at the start. However, as epochs increase, the validation accuracy becomes higher than training; in the end, it got 90% validation and 88% training accuracy.

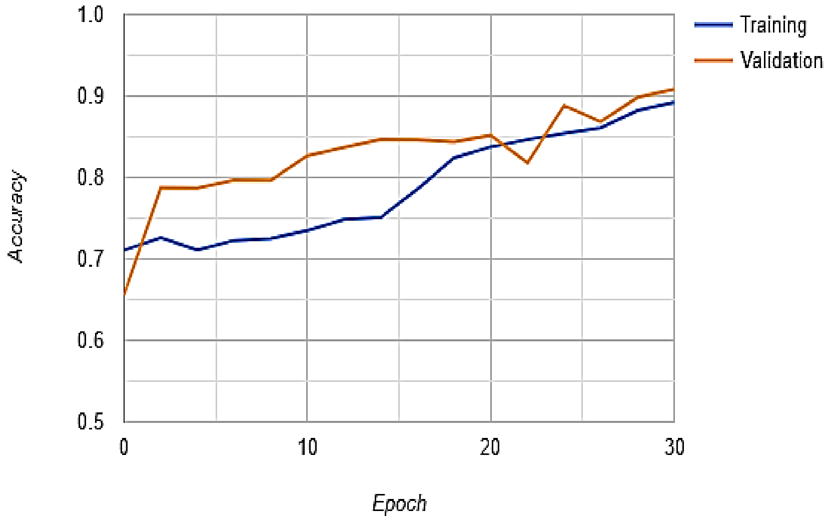


Figure 5: Proposed model accuracy

Compared to the suggested attention-based model, where the accuracy is about 91%, it achieves F-measure of 88% with pre-trained embedding. The comparison of both models is listed in Table 3.

Table 3: Comaprision of basline and proposed model

Model	**Pre trained word Embedding**		**No pre trained word embedding**	
	Accuracy	***F1-Score***	***Accuracy***	***F1-Score***
Baseline Model	86%	85%	84%	80%
Proposed Model	91%	88%	90%	87.3

A confusion matrix is a classifier performance assessment approach for machine and deep learning models. It is a table that shows how well a classifier performs on test data with known true values. Comparing actual

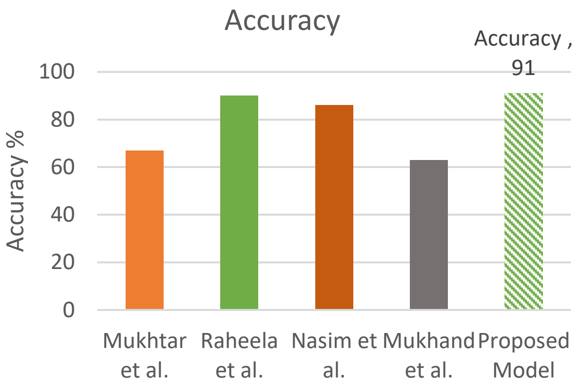


Figure 6: Accuracy of previous techniques

and anticipated classifications, the confusion matrix illustrates the quality of any classifiers. The proposed model is also evaluated using the confusion matrix metric. Figure 7 presents the confusion matrix for the proposed model. Various techniques have been used to classify Urdu sentiments in the last decade. Various authors used ML and DL techniques to classify various Urdu data sets. Table 4 presents the comparison of the proposed work with previous techniques.

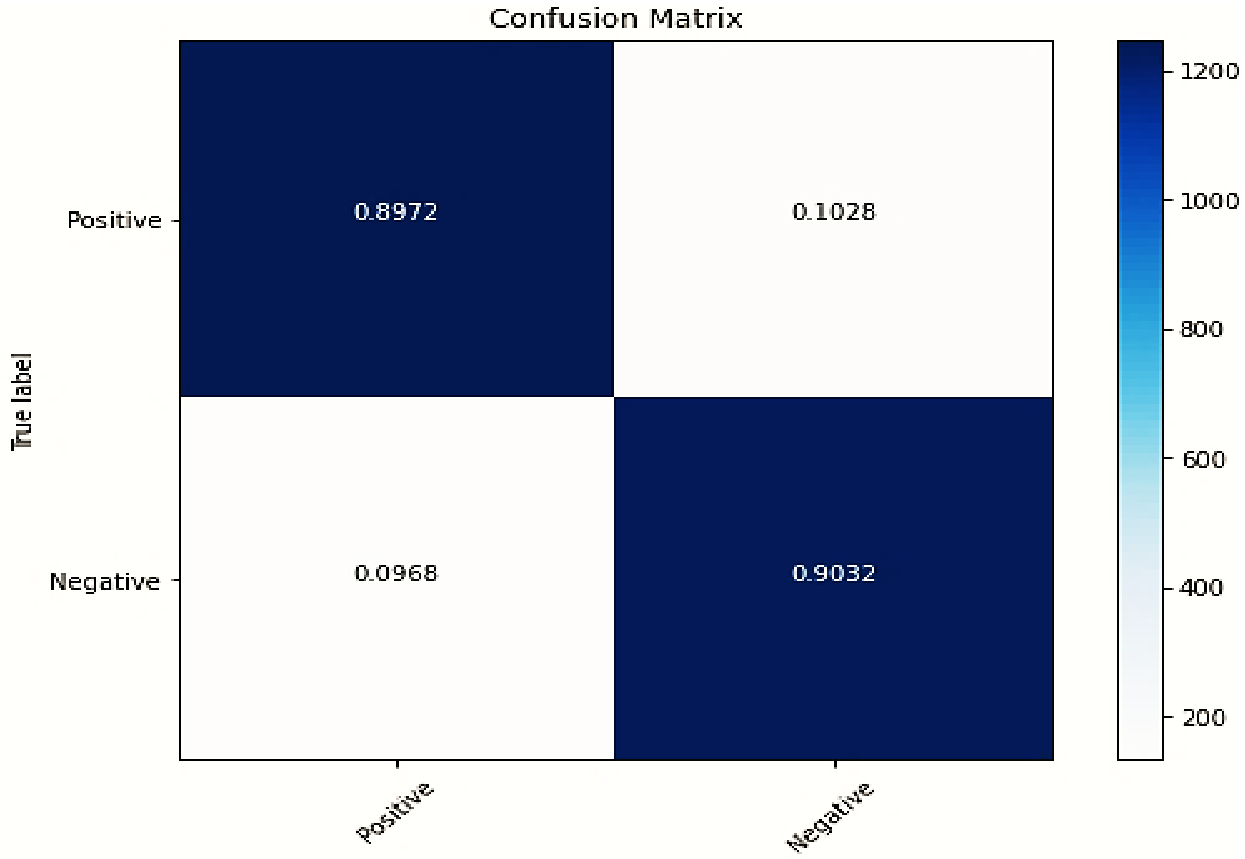


Figure 7: Confusion matrix for proposed model

Table 4: Comparison with previous techniques

Author	Year	Technique	Accuracy
Naqvi et al.	2021	LSTM, CNN, BiLSTM and attention based BiLSTM	77%
Hashim et al.	2016	Lexicon	86.8 %
Masood et al.	2022	LSTM	86.8 %
Safder et al	2021	SVM,CNN, LSTM, RCNN	84%
Mukhtar et al.	2016	SVM, KNN	67%
Proposed Model	**Proposed**	**Attention based Transfer Learning Model**	**91%**

7. Conclusion

Deep learning approaches have recently proven significant in various applications, including machine translation, image recognition, object detection, and natural language processing (NLP). The findings of this study's work are promising in Urdu SA. A deep learning model based on the attention mechanism is suggested in this work to categorize sentiment from Urdu text. First, preprocessing was done to normalize and tokenize the gathered data set. Secondly, a review-level sentiment classification algorithm is developed for analyzing the review's polarity by picking the most informative terms that indicate a specific sentiment in a review. The experiments were performed with the help of a transfer learning approach with pre-trained embeddings. Visualizing attention weights is also done to reveal the black box of the models and confirm their intuition, which helps interpret the model's acquired representations. The proposed system is evaluated based on accuracy and F1 score and compared with the baseline model. The findings show that the proposed deep learning model achieved 91% accuracy and 88% F1 score with pre-trained word embeddings. The model achieved 90% and 87.3% accuracy and F1-score, respectively, without pre-trained embeddings. In the future a large dataset by incorporating various Urdu dialects will get more reliable results. Moreover, the proposed technique can be trained on multilingual data for multilingual sentiment analysis.

References

Almani, N.M. and L.H. Tang. (2020). *Deep Attention-Based Review Level Sentiment Analysis for Arabic Reviews*, 6th Conference on Data Science and Machine Learning Applications (CDMA), IEEE.

Asghar, M.Z., et al. (2017). Sentence-level emotion detection framework using rule-based classification. *Cognitive Computation*, 868-894.

Bibi, R., et al. (2019). *Sentiment analysis for urdu news tweets using decision tree*, IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE.

Cambria, E., et al. (2017). *Affective computing and sentiment analysis*, in A practical guide to sentiment analysis, Springer, 1-10.

Daud, A., W. Khan, and D. Che. (2017). Urdu language processing: a survey, *Artificial Intelligence Review*, 279-311.

Glorot, X., A. Bordes, and Y. Bengio. (2011). *Deep sparse rectifier neural networks, in Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings.

Hashim, Faiza, and M. Khan. (2016). *Sentence level sentiment analysis using urdu nouns,* In Proceedings of the Conference on Language & Technology, vol. 2016.

Khalid, K., et al. (2017). *Extension of semantic based urdu linguistic resources using natural language processing*, IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and

Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE.

Khotimah, D.A.K. and R. Sarno. (2018). *Sentiment detection of comment titles in booking. com using probabilistic latent semantic analysis*, 6th International Conference on Information and Communication Technology (ICoICT), IEEE.

Kumar, H., B. Harish, and H. Darshan. (2019). Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method, *International Journal of Interactive Multimedia & Artificial Intelligence*.

Liu, H., Q. Yin, and W.Y. Wang. (2018). Towards explainable NLP: A generative explanation framework for text classification, *arXiv preprint* arXiv:1811.00196.

M. Masood, F. Azam, M. Waseem Anwar and J. Ur Rahman. (2022*). Deep-learning based framework for sentiment analysis in Urdu language*, 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), doi: 10.1109/ICoDT255437.2022.9787451, 1-7.

M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad. (2019). Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language, *Expert Systems*, e12397.

Mehmood, K., et al. (2019). Sentiment analysis for a resource poor language—Roman Urdu, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 1-15.

Mukhtar, N. and M.A. Khan. (2018). Urdu sentiment analysis using supervised machine learning approach, *International Journal of Pattern Recognition and Artificial Intelligence*, 1851001.

Mukhtar, N. and M.A. Khan. (2020). Effective lexicon-based approach for Urdu sentiment analysis, *Artificial Intelligence Review*, 2521-2548.

Mukund, S. and R.K. Srihari. (2012). *Analyzing Urdu social media for sentiments using transfer learning wit hcontrolled translations*, in Proceedings of the Second Workshop on Language in Social Media, 1-8.

Naqvi, U., A. Majid, and S.A. Abbas. (2021). UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods, *IEEE Access*, 114085-114094.

Nasim, Z. and S. Ghani. (2020). Sentiment Analysis on Urdu Tweets Using Markov Chains, *SN Computer Science*, 1-13.

Patel, J.M., Web. (2020). *Scraping in Python Using Beautiful Soup Library*, in Getting Structured Data from the Internet, Springer, 31-84.

Soleymani, M., et al. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 3-14.

Syed, A.Z., M. Aslam, and A.M. (2014). Martinez-Enriquez, Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text, *Artificial Intelligence Review*, 535-561.

Tan, C., et al. (2018). *A survey on deep transfer learning*, in International conference on artificial neural networks, Springer.

U. Naqvi, A. Majid and S. A. Abbas. (2021). UTSA: Urdu Text Sentiment Analysis Using Deep Learning Methods, *IEEE Access*, doi: 10.1109/ACCESS.2021.3104308, 114085-114094.

Yang, L., et al. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning, *IEEE Access*, 23522-23530.

Zhang, Z., D. Robinson, and J. Tepper. (2018). *Detecting hate speech on twitter using a convolution-gru based deep neural network*, in European semantic web conference, Springer.

Foundation University Journal of Engineering and Applied Sciences
Vol. 3, Issue 1, 2022.
DOI: 10.33897/fujeas.v3i1.571

Research Article

Effect of Preprocessing and No of Topics on Automated Topic Classification Performance

Ijaz Hussain[1*], Zafar Mehmood Khattak[2]

[1]Department of Computer and Information Sciences, PIEAS Islamabad, Pakistan.
[2]Department of Computer Science, University of Gujrat, Pakistan.
[*]Corresponding Author: Dr. Ijaz Hussain, Email: ijazhussain@pieas.edu.pk

Abstract:

The emergence of the Internet has caused an increasing generation of data. A high amount of the data is of textual form, which is highly unstructured. Almost every field i.e., business, engineering, medicine, and science can benefit from the textual data when knowledge is extracted. The knowledge extraction requires the extraction and recording of metadata on the unstructured text documents that constitute the textual data. This phenomenon is regarded as topic modeling. The resulting topics can ease searching, statistical characterization, and classification. Some well-known algorithms for topic modeling include Latent Dirichlet Allocation (LDA), Nonnegative Matrix Factorization (NMF), and Probabilistic Latent Semantic Analysis (PLSA). Different parameters can affect the performance of topic modeling. An interesting parameter could be the time required to perform topic modeling. The fact that time is affected by many factors applicable to topic modeling as well; however, measuring the time concerning some constraints can be beneficial to provide insight. In this paper, we alter some preprocessing steps and topics to study their impact on the time taken by the LDA and NMF topic models. In preprocessing, we limit our study by altering only the sampling and feature subset selection whereas in the second step we, have changed the number of topics. The results show a significant improvement in time.

Keywords: Text mining; Topic classification; Latent dirichlet allocation, Knowledge extraction

1. Introduction

The Internet has a direct influence on an increased amount of data being generated, collected, processed, and stored. The wide use of devices, particularly embedded devices, has increased the generation of data. Much of the data is in textual form but has the potential of having much more productive application than mere communication. Topic modeling algorithms have been designed to exploit the textual data by assigning topics to the constituent documents. This can lead to more intelligent searching, better statistical analysis of events, and hence better classification.

Topic modeling algorithms take as input some text documents and produce output as a set of topics, where each topic tends to describe the document it belongs to. Applications of topic modeling could be found in diverse areas, some of them include text-recommendation systems (Jin, Zhou, & Mobasher, 2005; Krestel, Fankhauser, & Nejdl, 2009), and digital image processing (Niebles, Wang, & Fei-Fei, 2008; Torralba, Willsky, Sudderth, & Freeman, 2005).

The variety of applications has led researchers to enhance the known topic modeling algorithms by proposing their variants as well as proposing novel algorithms (Naseem, Razzak, Eklund, & Applications, 2021). The requirement for a quick response from different software as well as hardware has led scientists to focus on optimized algorithm analysis, which has emphasized the requirement of producing quick algorithms. In the case of topic modeling, the time performance measure is highly important, for instance, the popularity of the Google search engine is particularly regarded as the best, depending on its high-quality results in reduced time.

Although it is known that the time required by any algorithm to perform its task can depend on many factors and thus explicitly expressed in terms of time seems to be discouraged by the scientific community, as they tend to express time as asymptotic notations rather than explicit units. However, due to the observation that significant work in topic modeling could be found implemented in python, it could be beneficial to build a quick vision regarding time using python functions. We have aimed to present an estimation of the time taken by LDA and NMF topic models by utilizing python programming language.

We study the impact of preprocessing on the time taken by the NMF and LDA topic models, as well as the effect of the number of topics. In preprocessing, we modify two steps, the first is "sampling"; selecting some part of data set objects as representative of the entire population. The second step is feature subset selection from the set of all features. In our case, we specify and change the sample numbers as well as the feature subset from all features using Scikit-learn.

The rest of the paper is organized as follows; Section 2 covers the background, Section 3 cites some related work, Section 4 performs a time-based comparison between NMF and LDA, Section 5 presents and discusses results, and the last but not least Section 6 gives the conclusions.

2. Background

We present some background of the basic concepts in chronological order, to help the reader grasp the concepts on which we build our analysis. The concepts are (i)Text mining, (ii) Topic modeling, (iii) Latent Dirichlet Allocation, and (iv) Non-negative Matrix Factorization.

2.1 Text Mining

Text mining refers to applying the data mining process to textual data (Naseem et al., 2021). It comprises the steps such as structuring the data sets, applying data mining tasks to find patterns in the structured dataset, and evaluating the results. Typical text mining tasks could include text categorization, text clustering, document summarization, keyword extraction, etc. In this research, we used the text mining utilities provided by Scikit-learn to exploit text-mining tasks on our selected dataset.

2.2. Topic Modeling

In the context of machine learning and natural language processing, topic models are generative models which provide a probabilistic framework (Ponweiser, 2012). They are generally used to automatically perform organizing, understanding, searching, and summarizing operations on large electronic archives. Topics obtained are used to predict variable relations between words in a vocabulary and their occurrence in a particular document. The significance here is that the relations are generally hidden, not estimated, and could be beneficial for important scientific and business needs. Topic models discover the hidden topics throughout a corpus and annotate the documents according to those topics. Each word is seen as drawn from one of those topics. Finally, a document coverage distribution of topics is generated and it provides a new way to explore the data from the perspective of topics.

Topic modeling aims to find the topics of a document. To develop insight, we can consider a simple example depicted in Figure 1, which illustrates the ideal case of the assignment of the topic to a given document. In Figure 1, the frequency of occurrence of each unique word is calculated, and then based on the topic of words with the highest frequencies, the topic of the document is decided. However, in practice, it is hard to

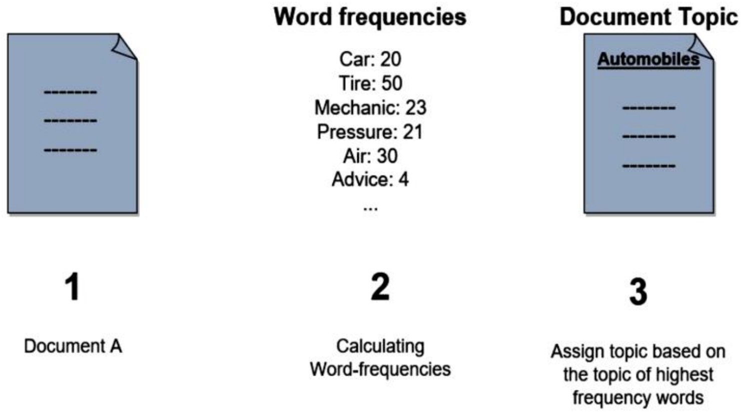


Figure 1: An ideal case of assigning the topics to a document

find exact topics and instead, approximation algorithms are required. LDA and PLSA are some significant topic modeling algorithms, with LDA being a gold standard. A study of generative models underlying PLSA and LDA reveals the assumptions that each topic is characterized by a specific word-usage probability distribution, and that every document in the corpus is generated from a topic mixture. For example, consider a corpus consisting of documents that are generated from two topics. Let the topics be mathematics and biology. If a document is highly concerned with biology, say it dbio, then it is much more likely for it to have p(topic = biology|dbio) = 0.9 and p(topic = mathematics|dbio) = 0.1.

It is possible to incorporate probability rules in topic models to enable them to predict topics, however, a particular topic model likely lacks in requisite time efficiency. The performance of a topic model has been a key of interest to researchers. Despite the high standard of LDA, researchers have tried to present novelty by targeting some performance measures. Time consumed by an algorithm is very vital when comparing algorithms, and the availability of topic model implementations has encouraged us to direct efforts toward evaluating different performance measures to select a topic model that suits our requirements.

2.3 Latent Dirichlet Allocation

LDA is a generative model that takes some data observations and generates some unobserved data to help in depicting the relationship of the taken observations (Grün & Hornik, 2011). In the context of topic-modeling, it works by representing the documents contained in a corpus, as a mixture of topics. A topic is a distribution over a fixed vocabulary. This topic mixture excludes the words with certain probabilities. These excluded words are likely to be irrelevant to the topic to be extracted. LDA has a huge influence in the fields of natural language processing and statistical machine learning and has become one of the most popular probabilistic text modeling techniques in machine learning.

LDA is capable of providing multiple topics (Jing, 2014). However, topics obtained by LDA potentially depend on the pre-processing, which makes it vital to apply pre-processing tasks, such as the removal of stop words. Applying LDA after pre-processing gives topics generated from the collection of documents (D. M. J. C. o. t. A. Blei, 2012). Over these topics, the corpus documents have some probability distribution. For example, a car topic could have words ”fuel”, and “engine” with high probability, and a topic on sports will have words like ”soccer”, and “cricket” with high probability.

From the generated probability distribution over the topics, each word is considered to be drawn from those topics. This contributes to knowledge about the frequency with which each topic is involved in a particular document.

2.4 Non-negative Matrix Factorization

NMF factorizes a given matrix, say X, into two matrices, say W and H. A restriction on these three matrices is that they must be non-negative, which is aimed to ease the particular analysis to be performed. However, numerical techniques are applied to approximate the problem as it is generally not exactly solvable. NMF can be exploited in text mining. For this purpose, the set of documents to be examined and various terms in it are taken and then assigned weights to them. From these weights, a document-term matrix is constructed, which is then factored into the term-feature and feature-document matrix. Finally, the feature-document matrix gives topics for data clusters of the documents.

3. Related Work

The bulk of data that we are currently assembling and storing is unprecedented. A challenge for its research is that approximately 80 percent of stored documents belong to unstructured text. As digital data preservation is increasing, there is a demanding need for fast and consistent algorithms to operate and convert into novel knowledge. One of the fundamental challenges in the arena of natural language processing is connecting the gap among information in text databases and their significance in particulars of its topics.

Topic modeling algorithms are significant for satisfying this breach. A database of text documents is used for topic models to automatically label individual documents in positions of the underlying topics (D. M. Blei & Lafferty, 2007), (Jin et al., 2005). K. Thilagavathi, and V. Shanmuga explore the current efforts and contributions in text mining techniques. Many data mining methods have been planned for mining appreciative patterns in text documents. Since most existing text mining approaches adopted term-based methods, they all suffer from the issues of polysemy and synonymy. An inventive and valuable pattern-finding technique that contains the processes of pattern deploying, to advance the efficiency of using and updating exposed patterns for finding appropriate and fascinating information, is discussed (Krestel et al., 2009). Probabilistic latent semantic analysis (PLSA) relies on fitting a reproductive model of the corpus. Explicitly, the model considered that a document in the corpus covers a combination of topics, and each topic is categorized by an exceptional word usage probability distribution. Text mining is an essential field of data mining that deals with unstructured data. It familiarizes several models from connected research areas like clustering, classification, etc.

Mathematical actions can be originated by applying Text analysis approaches to unstructured text material. The stemming technique produces a stem, which is a regular assembly of words with equivalent meanings. This technique labels the base of a specific word. Derivational and Inflectional stemming are different methods. One of the collective algorithms for stemming is porter's algorithm. For example, if a document relates the word resigned, resignation, and resigns as alike, then it will be deliberate as resign later applying the stemming technique (Thilagavathi & Shanmuga, 2014). As a significance, it better contests the statistical belongings of factual texts and explains many of the fundamental limitations of LDA. Fascinatingly, Topic Mapping provides only insignificant improvements concerning likelihood (because of the extraordinary degeneracy of the probability landscape) but produces much better accurateness and reproducibility (Lancichinetti et al., 2015; Monali, Sandip, & Engineering, 2014).

Outdated collaborative filtering to digital reproducing is challenging because manipulation of facts is very sparse due to the unusual volume of documents compared to the number of users. Content-based methodologies, on the other hand, are smart because textual content is very enlightening.

In large-scale content-based cooperative filtering for digital dissemination, to decipher the digital publishing recommender difficulty, two approaches are associated: LDA and deep belief nets (DBNs) that equally find low-dimensional latent illustrations for documents. Well-organized retrieval can be supported in the latent representation (Arora et al., 2013). In LDA, the parameterization of topics is done as categorical deliveries over impervious word categories with multivariate Gaussian scatterings on the implanting space. This stimulates the model to cluster words that are a priori recognized to be semantically associated with topics.

To accomplish interpretation, a fast warped Gibbs sampling algorithm built on Cholesky decompositions of covariance atmospheres of the subsequent predictive distributions is presented in (Arora et al., 2013). Further originates a scalable algorithm that draws examples from predictive distributions and fixes them through a Metropolis-Hastings phase (Arora et al., 2013). Both LDA and DBN methods trust on the expansion of a likelihood that hangs on non-linearly on a huge amount of variables, an NP-hard problem. Even though it is fine known that the difficulty is computationally tough, little is recognized about how, in preparation, the roughness of the possibility landscape influences an algorithm's performance. To increase a more thorough theoretical consideration, and implementation of an organized examination of a highly identified and built data set is presented in (Gruber, Rosen-Zvi, & Weiss, 2012). This high point of control permits to tease separately the theoretical restrictions of the algorithms from further causes of error that would be generally uncontrolled in traditional datasets. The investigation exposes that the standard methods for likelihood optimization are delayed by the very irregular topology of the countryside, even in very modest cases such as when topics practice exclusive vocabularies. In (Gruber et al., 2012), the authors illustrate that a networking attitude to topic modeling empowers pointing to the likelihood landscape more efficiently, and produce supplementary accurate and reproducible consequences.

4. Methodology

To compare the performance of NMF and LDA with respect to time, we utilize the programming functionality provided by Python as well as the topic models from Scikit-learn repositories (Sontag & Roy, 2011). From Sikit-learn repositories, we build our model on the work done by O. Grisel, L. Buitinck, and C. Yau (Jung, Shin, & Lee, 2022) that performs topic modeling on the 20 newsgroups dataset. Before performing a comparison, we present the details of some significant preprocessing steps taken from (Jung et al., 2022):

a. Sampling: Default samples are 2000. The number of samples is alterable and is stored in a variable.
b. Feature subset selection: Default features are 1000. The number of features is alterable and is stored in a variable.
c. Variable transformation: The textual data is transformed into numeric data. For LDA, the term frequency (tf) is applied, for which the collection of text documents is converted into a matrix of token counts, which is then used to produce the term-document matrix. Whereas, for NMF the term frequency-inverse document frequency (tf-idf) is used to convert the text document collection into a term-document matrix. After preprocessing, we apply LDA and NMF to form some conclusions about the time-based comparison between LDA and NMF, we conduct five experiments for each of the following criteria:

i. Number of topics.

ii. Number of samples.
iii. Number of features.

5. Results and Discussion

The results are composed of multiple runs of the topic modeling implementation for each of the mentioned criteria (i.e., number of topics, samples, and features). We restrict the runs to five runs per criterion, where each run has an altered criterion value.

5.1 Effect of Number of Topics on Time

The impact of the number of topics on the time taken by LDA and NMF is studied, while the number of samples and number of features is set to 2000 and 1000, respectively. We start from the number of topics set to 10 and alter it by the increment of 10 units till we reach 50. The experiment results are shown in Table 1, where ten topics takes about ten seconds using LDA approach and 50 topics take about fifteen seconds. To choose the optimal number of topics we can use validation perplexity that is low on greater number of topics.

Table 1:Comparison of NMF and LDA with respect to time and no of topics

Sr. No	Number of topics	LDA Time (seconds)	NMF Time (seconds)
1	10	9.869	1.052
2	20	10.298	1.957
3	30	12.598	2.427
4	40	12.720	3.206
5	50	14.049	4.064

5.2 Effect of Number of Samples on Time

In this experiment we study, the impact of number of samples on the time taken by LDA and NMF, while the number of topics and the number of features are set to 10 and 1000, respectively. We start from the number of samples set to 500 and alter it by the increment of 500 units till we reach 2500. The experiment results are shown in Table 2.

Table 2:Comparison of NMF and LDA with respect to time and no of samples

Sr. No	Number of topics	LDA Time (seconds)	NMF Time (seconds)
1	500	3.112	0.237
2	1000	5.599	0.446
3	1500	9.423	0.882
4	2000	13.239	1.141
5	2500	14.290	0.707

5.3 Effect of Number of Features on Time

In this part of the experiment we study the impact of number of features on the time taken by LDA and NMF, while the number of topics and the number of samples are set to 10 and 2000, respectively. We start from the number of features set to 200 and vary it by the increment of 200 units till we reach 1000. The experiment results are shown in Table 3.

Table 3:Comparison of NMF and LDA with respect to time and no of features

Sr. No	Number of topics	LDA Time (seconds)	NMF Time (seconds)
1	200	10.162	0.606
2	400	12.175	0.839
3	600	11.302	1.333
4	800	10.491	0.818
5	1000	9.061	1.111

5.4 Average Performance for the Criteria

An analysis of the average time taken by NMF and LDA is presented in Table 4. In the case of the topic criterion, the time taken by LDA is 11.777 seconds, and that of NMF is 1.9418 seconds. The time taken by LDA on average is approximately six times greater than the time taken by NMF. In case of the sample criterion, the time taken by LDA is 9.1326 seconds, and that of NMF is 0.6826 seconds. The time taken by LDA on average is greater than 13 times the time taken by NMF. In case of the feature criterion, the time taken by LDA is 10.6382 seconds, and that of NMF is 0.9414 seconds. The time taken by LDA on average is greater than 11 times the time taken by NMF.

Table 4:Comparison of NMF and LDA with respect to three criteria

Sr. No	Criterion	Average LDA Time (seconds)	Average NMF Time (seconds)
1	Topics	11.777	1.9418
2	Samples	9.1326	0.6826
3	Features	10.6382	0.9414

6. Conclusions

In all the three cases (i.e., topics, samples, and features) the time taken by NMF has been significantly less than the time taken by LDA. Therefore, we deduce that NMF tends to outperform LDA with respect to time. Considering the behavior with respect to the three criteria, we deduce that LDA tends to be most affected by altering the number of samples; as in case of samples is depicting increasing time for the five iterations. In case of altering the number of topics, the Table 2 also shows an increasing tendency of time, however, the depiction in case of 30 topics suggest that LDA might not follow a significant increasing pattern with respect to number of topics. In case of altering the number of features (Table 3), it might be deduced from the table that increase in features from 400 to 1000 LDA tends to take less time when a specific number of features is exceeded, which in this case is 800. Concluding the performance, LDA could perform better when if we take small samples, and might improve it by reducing number of topics to be predicted and increasing the number of features beyond a threshold.

In the case of NMF, it tends to take least time on average when we vary the number of samples. The effect on average time might not be evident from the graphs but taking the average of time taken for each of the three criterion (i.e., number of topics, samples and features) we get the least value of 0.6826 seconds for sample criterion as compared with the average time of 0.9414 and 1.9418 seconds for features and topic criterion, respectively. In case of topic criterion, although the time (Table 2) shows an increasing tendency, it is difficult to deduce due to the decrease in time at 20 and 50 topics. In case of sample criterion, the time (Table 3) shows an increasing tendency, with only exception at the last value of 2500. In case of features criterion, although the time (Table. 4) depicts an increasing tendency, but the values at 800 and 1000 tends to make it difficult to judge the behavior with the limited observations. Concluding the performance of NMF, it might be feasible to increase sample size but refrain from increasing number of features, and especially the number of topics, when a quick time-performance is required.

References

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D.,Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. Paper presented at the International conference on machine learning.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science.

Blei, D. M. J. C. o. t. A. (2012). Probabilistic topic models. 55(4), 77-84.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. J. a. p. a. (2012). Latent topic models for hypertext.

Grün, B., & Hornik, K. J. J. o. s. s. (2011). topicmodels: An R package for fitting topic models. 40, 1-30.

Jin, X., Zhou, Y., & Mobasher, B. (2005). A maximum entropy web recommendation system: combining collaborative and content features. Paper presented at the Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.

Jing, Q. (2014). Searching for economic effects of user specified events based on topic modelling and event reference. Acadia University,

Jung, G., Shin, J., & Lee, S. J. A. I. (2022). Impact of preprocessing and word embedding on extreme multi-label patent classification tasks. 1-16.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. Paper presented at the Proceedings of the third ACM conference on Recommender systems.

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. J. P. R. X. (2015). High-reproducibility and high-accuracy method for automated topic classification. 5(1), 011007.

Monali, P., Sandip, K. J. I. J. o. A. R. i. C., & Engineering, C. (2014). A concise survey on text data mining. 3(9), 8040-8043.

Naseem, U., Razzak, I., Eklund, P. W. J. M. T., & Applications. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. 80, 35239-35266.

Niebles, J. C., Wang, H., & Fei-Fei, L. J. I. j. o. c. v. (2008). Unsupervised learning of human action categories using spatial-temporal words. 79, 299-318.

Ponweiser, M. (2012). Latent Dirichlet allocation in R.

Sontag, D., & Roy, D. J. A. i. n. i. p. s. (2011). Complexity of inference in latent dirichlet allocation. 24.

Thilagavathi, K., & Shanmuga, V. J. I. J. A. R. C. S. R. (2014). A survey on text mining techniques. 2(10), 41-50.

Torralba, A., Willsky, A., Sudderth, E., & Freeman, W. J. A. i. n. i. p. s. (2005). Describing visual scenes using transformed dirichlet processes. *18*.

Foundation University Journal of
Engineering and Applied Sciences
Vol. 3, Issue 1, 2022.
DOI: 10.33897/fujeas.v3i1.381

Research Article

Nanotechnologies: AI Weapons Governing the Military Battle Field

Atif Ali[1*], Zulqarnain Fareed[2], Shaikh Muhammed Nadeem[2], Hina Naseem[3], Khushboo Farid Khan Ghouri[2], Muhammad Shareh Qazi[4]
Muhammad Faisal Khan[5]

[1]Research Management Centre (RMC), Multimedia University, Cyberjaya 63100 Malaysia
[2]University of Karachi, Pakistan
[3]Allama Iqbal Open University, Islamabad, Pakistan
[4]University of the Punjab, Lahore
[5]Riphah International University, Islamabad, Pakistan
*Corresponding Author: Atif Ali. Email: atif.alii@yahoo.com

Abstract:

With major advantages and concerns, nanotechnology (NT) is expected to change several industries. New risks could emerge due to military advancements, requiring additional planning and work to contain them. When it comes to military R&D, the NT is moving forward quickly. Future uses could benefit all military branches. Microrobots and new biological weapons could endanger stability and arms control. Many people are interested in nanotechnology as a scientific subject because of all the opportunities it offers. Nanorobots can be employed in various fields, including materials science, space exploration, ecology, information technology, electronics, and communications. On the other hand, these novel uses for nanorobotics in military applications and armament are revolutionary. An essay on the most recent developments in military nanorobot applications has been made available. Due to its fundamentally revolutionary advantages, military nanotechnologies have been argued to be more lethal than nuclear weapons for the entire planet and capable of being used in all conflict zones.

Keywords: Artificial Intelligence; Nanotechnology (NT); Grey goo; Robotics; Nanosat; Nano-robots

1. Introduction

Before today, there had been very little scholarly research on military NT. Government papers, conferences, military periodicals, and the general public have all disputed the topic. Once the data has been translated and sorted into combinations of 1 and 0 on a computer, it may be easily duplicated and transferred. Atoms, the basic constituents of matter, are the starting point for forming molecules. Nanotechnology allows these atoms and molecules to be modified quickly and economically. Today, every known chemical may be synthesised, replicated, and utilised (Ali, Qasim, et al., 2022). People benefit from this.

"Technology" will soon devolve into "nanotechnology, and we will lose touch with it. Engineers and scientists can use nanostructured devices to exploit their distinctive qualities. As material dimensions are shrunk to the nanoscale, new qualities are found and controlled in nanotechnology (Ali, Hafeez, Hussain, et al., 2020). At this extremely small scale, quantum mechanics succeeds Newtonian physics, leading to extraordinary material transformations. Scientists and engineers in the military believe that they must look into the implications of nanotechnology and use what they learn to safeguard their populations (Ali, Qasim, et al., 2022), (Ali, Hafeez, Hussain, et al., 2020).

2. Related work

2.1 Nanotechnology in the Military

The DoD is working on developing compact, more potent bombs than are now available, made of chemical explosives with ultra-high burn rates. A metastable intermolecular molecule is called a nano-thermite (MIC). As their name suggests, nano weapons use nanotechnology to improve existing military capabilities. Research and development for "mini-nuke" devices are underway in the US, Russia, and Germany. The ability to restrict mass destruction weapons is being pushed to its limit by the development of smaller nuclear bombs. According to researcher and CBRNE specialist Andy Oppenheimer (chemical, biological, nuclear, radiological, and explosive weapons), Oppenheimer continues, "[The bombs] are capable of destroying everything. Every threat grows. (Thomas et al., 2021).

Other nations compete for this military nanotechnology material, including the US and Europe, to alter their military strategies, such as MEMS (microelectronic mechanical systems) (as shown in Figure 1). Microelectronic mechanical and micrometric intelligence devices will guide this mechanism to its intended spot.

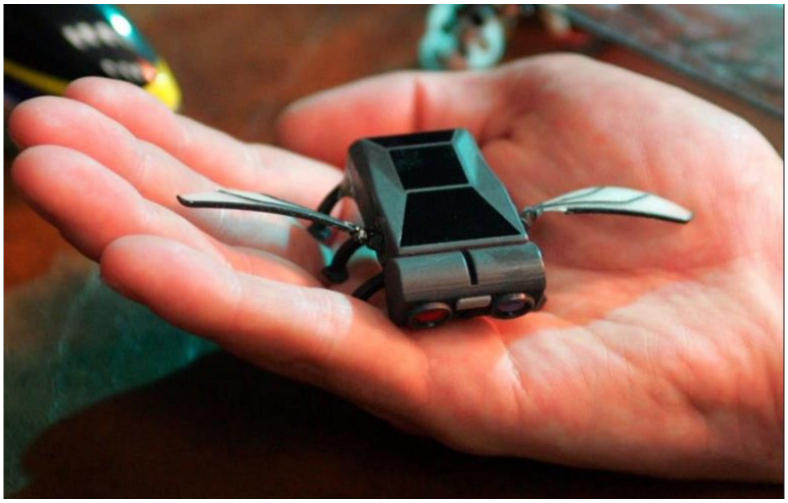


Figure 1: Microelectronic Mechanical Systems (MEMS)

The development of global intelligence and MEMS-based terrorism detection systems is the focus of current research (Ali, Hafeez, Hussainn, et al., 2020),(Huang et al., 2021). These gadgets will create a "network" with external supercomputers, aided by computer programme optimizations, and include computers giving orders to the same equipment, confirming the upcoming battles that Europe and the US are witnessing. They will be dropped using a sensor-equipped projectile that explodes harmlessly in the designated location (Ali, Hafeez, Hussainn, et al., 2020). These surveillance devices will be almost invisible, raising significant ethical and legal questions.

The military's financing arm is commissioning the next generation of disaster-response robots for mad scientists, DARPA, as part of a new award scheme. The most notable aspect of it will be its size. A SHRIMP initiative (Short-Range Independent Micro Robotic Platforms) will concentrate on tiny robots that can fit through a garden hose or down a drainpipe. Dr. Ronald Polcawich, a DARPA project manager at the Microsystems Technology Office (MTO), remarked that robots "can give much-needed assistance and support in a natural disaster scenario, a search-and-rescue effort, or another urgent relief requirement" (Huang et al., 2021). A sample is shown in Figure 2.

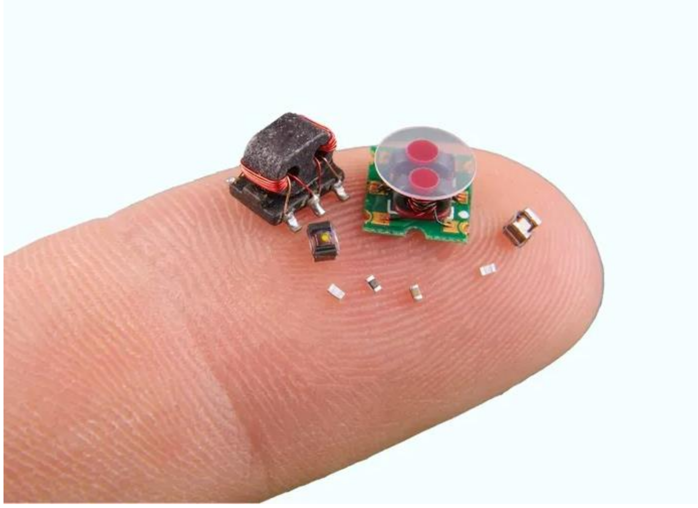


Figure 2: Independent Micro SHRIMP (short for SHort-Range robotic Platforms)

"Larger robotic platforms, on the other hand, cannot explore a variety of locations. Smaller robotics systems could be a big help, but downsizing these platforms requires more progress in the underlying technology."

2.2 Nano Drones

Rapid advancements in nano-UAV technology have produced new agility capabilities for national security requirements. UAS technology has improved recently, enabling them to fly farther and faster and carry out more challenging surveillance missions. The nano drone helped field employees understand local situational awareness. They are fairly heavy and small enough to carry in one hand. Figure 3 (Caliskan & Sokullu, 2019) depicts the Black Hornet Personal Identification System, the smallest combat-tested nano drone in the world.

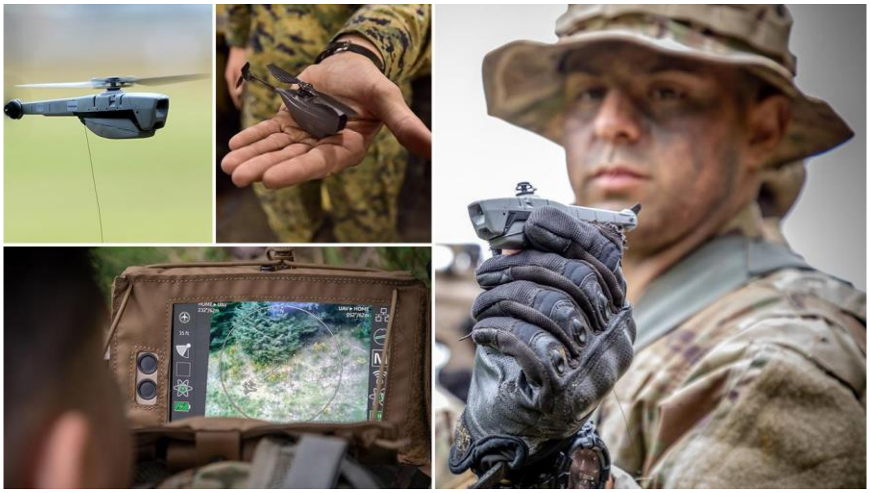


Figure 3: Black Hornet Personal Identification System (the tiniest nano drone ever tested in battle)

2.3 The use of Nanotechnology by Soldiers

This technique creates strong, durable, sensory, and active materials.

- *Nano-Armor:* Tungsten, rather than carbon, is used as the fundamental material in another process for making strong materials.
- *Lightweight Protective Clothes:* Antiballistic fabrics that decontaminate themselves, as well as nanofiber fabrics that disinfect themselves
- *Auxiliary Supports* Exoskeletons and robotics to help manned jobs and flexible/rigid fabrics for added strength.
- *Adaptive Suit:* Microsensors for ambient and situational awareness, brain and body sensing, integration into a smart suit or helmet, wearable and flexible screens for visual input, and switchable fabric for improved temperature control.
- *Smart Helmet:* The future soldier's fighting equipment will include a smart helmet in significant quantities. This smart helmet consists of a helmet with an intelligent multi-sensor system and serves as a platform system for many purposes (shown in Figure 4) (Stoudt, 2012):
    - Low-directed power, efficient communication, and RF array antennas for locating friends, RFID, and low-directed power.
    - Arrays of speakers (microphones).
    - B/C sensor arrays, an optical/IR camera with 360-degree vision, and early warning systems.
    - EEG wireless sensor.

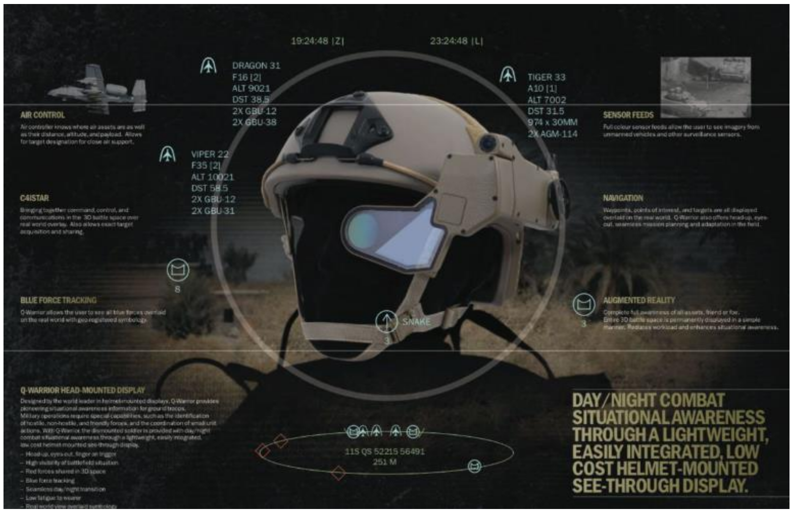

19:24:48 |Z|
23:24:48 |L|
DRAGON 31
F16 [2]
ALT 9021
DST 38.5
2X GBU-12
2X GBU-38
TIGER 33
A10 [1]
ALT 7002
DST 31.5
974 x 30MM
2X AGM-114
VIPER 22
F35 [2]
ALT 10021
DST 58.5
2X GBU-12
2X GBU-31
AIR CONTROL
Air controller knows where air assets are as well as their distance, altitude, and payload. Allows for target designation for close air support.
C4ISTAR
Bringing together command, control, and communications in the 3D battle space over real world overlay. Also allows exact target acquisition and sharing.
BLUE FORCE TRACKING
Q-Warrior allows the user to see all blue forces overlaid on the real world with geo-registered symbology.
Q-WARRIOR HEAD-MOUNTED DISPLAY
SENSOR FEEDS
Full colour sensor feeds allow the user to see imagery from unmanned vehicles and other surveillance sensors.
NAVIGATION
Waypoints, points of interest, and targets are all displayed overlaid on the real world. Q-Warrior also offers head-up, eyes-out, seamless mission planning and adaptation in the field.
AUGMENTED REALITY
Complete full awareness of all assets, friend or foe. Entire 3D battle space is permanently displayed in a simple manner. Reduces workload and enhances situational awareness.
8
SNAKE
3
3
11S QS 52215 56491
251 M
DAY/NIGHT COMBAT SITUATIONAL AWARENESS THROUGH A LIGHTWEIGHT, EASILY INTEGRATED, LOW COST HELMET-MOUNTED SEE-THROUGH DISPLAY.

Figure 4: Smart helmet

Utilizing microsystems and nanotechnology, it will be possible to reduce the weight of the apparatus currently fastened to or placed on the helmet, relieving some of the strain on the head (Jesudoss et al., 2019).

2.4 Arms with High Capacity

Nanotechnology will be used to create next-generation weapons and dramatically boost the destructive capability of already-existing weapons. For instance, the weapons will be so potent that they may still hit their target despite DNA readings. It would be substantially more difficult for radar to identify aviation equipment because it would be produced with the least amount of metal possible, making it lighter and more efficient.

Weapons based on nanotechnology may be more deadly than those that are nuclear, chemical, or biological. Because any government may beat its foe in the first attack without worrying about retaliation, nuclear deterrence will be useless. For instance, a plane dumping nanorobots on an adversary's territory may destroy electronic equipment, sneak up on soldiers, and sleep in their blood until activated. These are a few of the scenarios used by military strategists. Fundamentally, terrorist organisations and small nations will have easier access to nanotechnological weapons than conventional ones. Due to the many uses for nanotechnology in society, materials will be broadly accessible.Other nations, including the US, have reportedly seen these weapons (Ali, 2021).

2.5 Nanosatellites

The field of space exploration is where nanotechnology has the most immediate applicability. We can talk about space stations, light, incredibly durable vehicles, personal spaceships, high-altitude launch facilities, and well-known nanosatellites such as the NANOSAT, a Spanish nanosatellite project that began in 1995.

The INAT (National Institute of Aerospace Technology) designed the NANOSAT, which is managed and built in Spain and based on a new design philosophy: smaller, more powerful, faster, with a specialised use, more benefits, and lower consumption. Spain will likely lead the "little revolution in space" (Figure 5) due to the success of this avant-garde project (You, 2018).

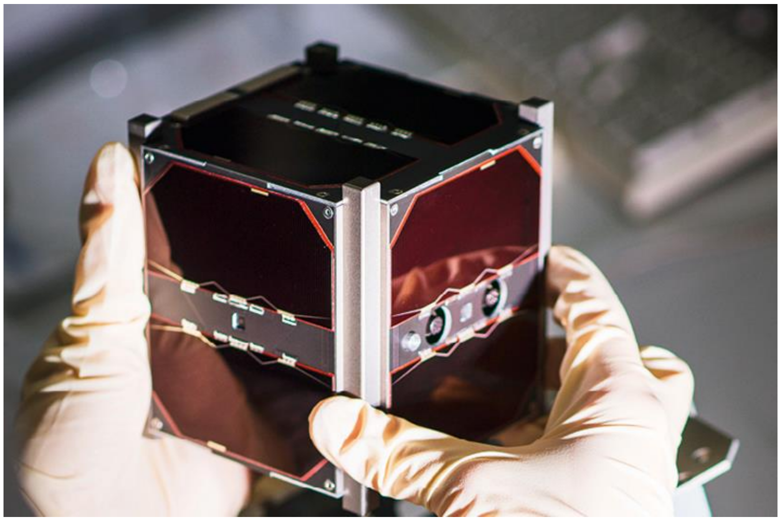


Figure 5: Nano peruvian satellite

2.6 Nanosatellites with a Low Orbit

Nanostructures will help the military see "the top of the hill" and improve communication.The Army Command and Missile Defense Force launched the first satellite impact nanosatellite nanotubes (SMDC-ONE). SMDC senior scientist Travis Taylor described it as a space cell phone tower for Army radios (Ali, Said, et al., 2022).

To provide soldiers stationed in remote areas with wifi connections, they plan to launch tiny satellites into

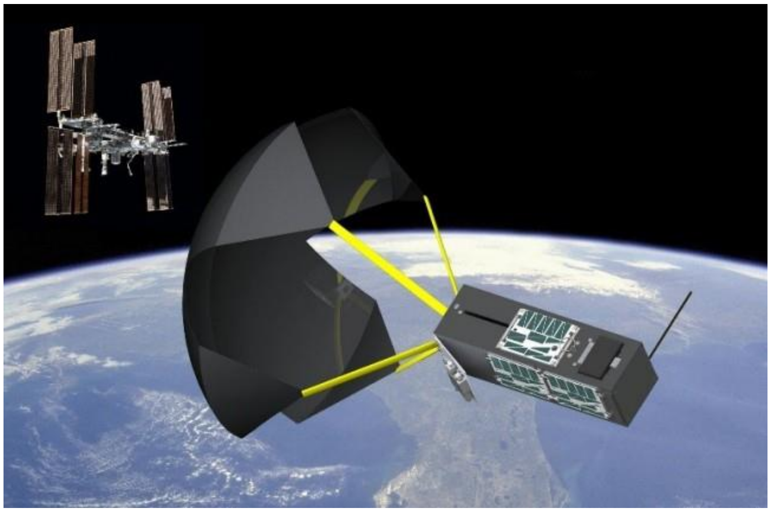


Figure 6. Small satellites in low earth orbit LEO

low earth orbit (LEO), 1,200 miles above the Earth. Since they are 60 times closer to the Earth than geo-synchronized communication satellites, low-earth orbit (LEO) microsatellites may transmit spotty signals using handheld radios (Figure 6).

2.7 Nano Missile

The RS-24 missile will enter service following the expiration of the Start disarmament pact on December 5, according to the Russian army, which also claims that a new intercontinental missile with nuclear warheads will be deployed by the end of the year. Live news is available on Interfax. Russia is worried about having to park missiles with multiple explosive heads due to the pact's expiration. In NATO circles, the parking comment is viewed as a regular upgrade phase because Russia desperately needs to update its nuclear arsenal of missiles from the Soviet era. The Russian Military Complex successfully uses nanotechnology to produce weapons and other military supplies. It has been announced that the Russian Military Complex's operations will be funded by this area of research, which has the potential to alter the dynamics of combat significantly. The director of the Nanotechnology Center at the Moscow Energy Institute claims that this speciality can "intelligently" strike moving targets like war machines. The clouds produced by these sub-millimetre devices have the potential to be destructive and of any magnitude. Moscow has reportedly invested more than 1.1 billion dollars in the advancement of nanotechnology, claims a Russian news site (You, 2018). Moscow and Washington want to discuss a new treaty to replace the Strategic Arms Reduction Treaty. The 1994 agreement, effective in 1994, places a 6,000 missile and 1,600 carrier weapons cap on each nation's strategic arsenal. (Debnath, 2016).

2.8 Gray goo

Grey goo is a nightmare nanotechnology scenario in which out-of-control self-duplicating nanobots destroy the biosphere by continually replicating themselves and gorging on life-sustaining elements. Consider a robot that floats in a soda bottle and is too small to be seen. Copying takes less than a minute. In the next minute, the two robots build two more. The globe will be converted into a large and obnoxious robotic ocean shortly. The main goal of these thieves is to replicate; to do so, they'll need fuel, which means they'll consume everything on the way and anything else, as seen in Figure 7 (Fries, 2018).

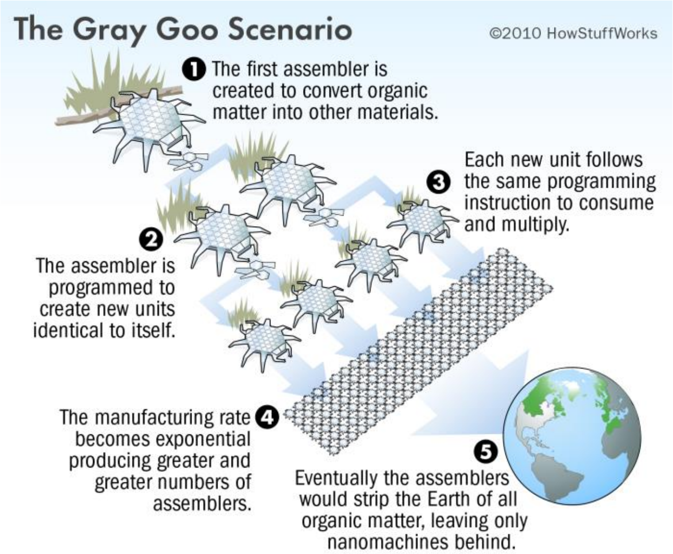

The Gray Goo Scenario
©2010 HowStuffWorks
1 The first assembler is created to convert organic matter into other materials.
2 The assembler is programmed to create new units identical to itself.
3 Each new unit follows the same programming instruction to consume and multiply.
4 The manufacturing rate becomes exponential producing greater and greater numbers of assemblers.
5 Eventually the assemblers would strip the Earth of all organic matter, leaving only nanomachines behind.

Figure 7.Gray Goo's nightmare

Is molecular nanotechnology capable of creating yet another doomsday scenario in which unchecked self-replicating robots devour all life on Earth while creating more of themselves, a behaviour known as ecophagy (literally, "eating the habitation")? The original theory was that this capability was built into computers, although popularizations suggested it may happen accidentally. Mathematician John von Neumann created the macroscopic self-replicating machines known as von Neumann machines, often known as clanking replicators. Mr Eric Drexler, a pioneer in nanotechnology, coined the phrase "grey goo" in his 1986 book "Engines of Creation." He said, "I regret not having created the phrase grey goo," in 2004. Although "grey goo" is mentioned twice and discussed in Engines of Creation, the term was first made public in November 1986 in Omni's mass-circulation magazine (Baber, 2004).

2.9 Genocide Weapons based on Nanotechnology

Nanoweaponry is plausible given the current rate of technological improvement, the convergence of genetic engineering, nanotechnology, and robots, and the fact that some people consider it science fantasy (GNR). Bioengineered viruses, self-replicating nanobots, and other innovative technologies and deployment tactics pose a potentially lethal threat (Goyal et al., 2013). In the future, nanobots will be utilised for genocide.

A potential ethnic bioweapon (biogenetic weapon, Figure 8) is a bioweapon that is designed to target people with particular genotypes or ethnicities.

2.10 Brain nanobots

In the 2030s, doctors will implant nanobots into the brains of live people to retrieve memories of those who have died (Figure 9).

By 2030, according to Kurzweil, our brains will be sufficiently powerful to connect to the cloud, enabling us to receive emails and photos directly in our heads and create backups of our memories and ideas. He asserts that it would be impossible to prevent if nanobots were microscopic robots made of DNA strands and floating inside the capillaries of our brains. Similar to how our ancestors learnt to use tools, he views the extension of our brain into mostly nonbiological thinking as the next stage in human development.

Figure 8: Racial bioweapon (biogenetic weapon)

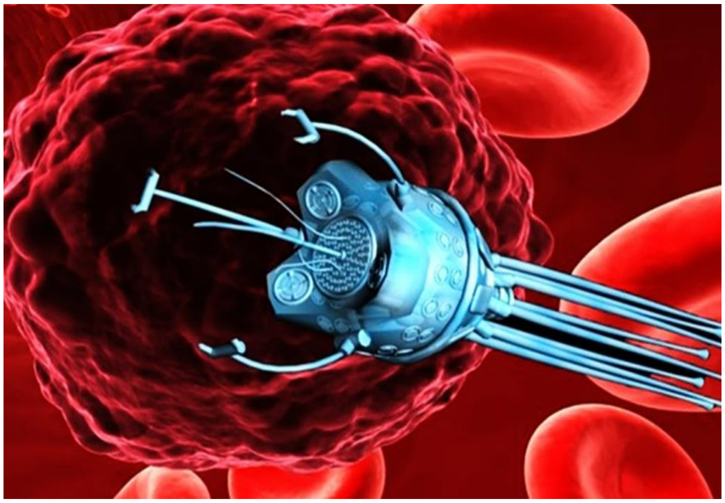


Figure 9: Nanorobots used for a human brain

And he asserts that this growth will raise our emotional and intellectual intelligence. To illustrate, he created a fictitious scenario with Larry Page, a co-founder of Google (Kita & Dobashi, 2015)]. He said, "We're going to construct detailed expression levels and add more levels to the hierarchy of brain modules.

3. Advantages and Disadvantages of Nanotechnology

Nanoparticles may harm biological systems and the environment due to free radical toxicity, which can damage DNA and lipids. Negative implications include economic disruption and potential threats to safety, privacy, health, and the environment (Ali, Ghouri, et al., 2022). Rapidly predicting the environmental impact of nanoparticles is essential in this context.

Nanotechnology is being developed for new instruments to replace a wide range of cellular machinery. Military research, such as reproductive science and technology, might be accelerated by nanotechnology, enabling the production of many weapons. Nanoscale devices may be utilized to make agricultural systems smarter in the future. Nanotechnology has several potential advantages in the military but has certain

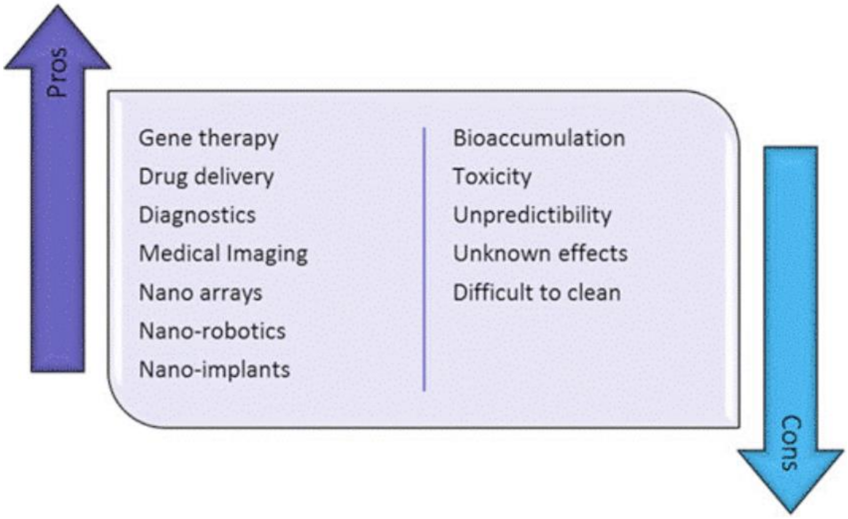


Figure 10: Pros and cons of nanotechnology

disadvantages. It's impossible to establish whether nanotechnologies are entirely safe or potentially harmful to human health. The effects of long-term exposure to nanomaterials, their unidentified life cycles, their interactions with biotic or abiotic environments, and possibly increased bioaccumulation have not been discussed. Before these applications move from the lab to the field, these factors must be considered. Commercializing nanotechnology is frequently hampered by high processing costs, R&D scaling restrictions for prototype and industrial production, and public perception issues involving environmental, health, and safety risks. Governments should set uniform, strict standards and oversight before commercialising and broadly using these nanotechnologies.

4. Nanotechnology Potential Threats

Humans gain from technology, but it also presents a risk. Ulrich Beck, a German scholar, suggested in Risk Society. The "risk society notion" has sparked a growth in research. After extensive research, he concluded that "we can immediately assume that society's technical dangers have likewise increased" because the income increase is proportional to the increase in social risk. Because of this, people are more curious about nanotechnology than genetic technology. Since publishing a nanotechnology development strategy in 2000, the United States has argued that the dangers of nanotechnology should be researched. Since 2001, the National Science Foundation (NSF) in the United States has investigated the socioeconomic effects and safety implications of nanotechnology-as-a-service. The Environmental Protection Agency has identified nanomaterials threatening human health and the environment (EPA). In 2005, (Fujitsuna, 2019) the US government developed the Nano, Environmental, Health, and Safety Strategic Research Program to protect public health. This curriculum focuses on risk management and scientific risk assessment.

Kanghe Environment is a non-profit organisation that conducts risk assessments for nanotechnology and promotes the development of innovative technologies for the benefit of society. In December, the US government's OECD sponsored a conference on the "Safety of Artificial Nanomaterials" in Washington, DC. In Europe, the United States, and Japan, twelve scientific and management symposiums on the effects of nanoparticles on biology and the environment were organised in 2006. In 2009, the Hugh Stan Nano Health Alliance and the Food and Drug Administration teamed up to study the behaviour and effects of nanoparticles on biological systems. Japan sponsored a conference on the toxicity of synthetic nanoparticles in 1990. The Nanotechnology Policy Advisory Committee was established in 2005 with the objectives of generating nanomaterials, reviewing the programme, and researching a systematic technique for evaluating the safety of nanoparticles. Japan requested health data from rice companies in March 2009. The European Commission and Royal Society have published nanomaterial regulatory issues since 2008. These documents regulate and guide nanotechnology research and social governance (Maguire, 2007). An EU nanotoxicology and safety research initiative promotes collaboration. Between 2005 and 2011, the EU supported 24 prominent programmers. Global nanotechnology adoption requires sensible nanotechnology production. Nanotechnology safety and ethics are debated in China. "Nanomaterials Technical Standards" were introduced in April 2005 (Spece et al., 2014).

The Chinese Academy of Sciences Institute of High Energy Physics founded the "Nano Biological Effect and Safety Joint Lab" in 2006. Since 2005, the Chinese National Natural Science Foundation and 973 research projects have funded nanotechnology safety studies (Olejníček et al., 2018).

5. Discussion

Even if molecular NT is still a long way off, military NT applications, as detailed above, have many potential dangers and issues. As a result, there are compelling reasons to investigate the problems and seek preventative steps as soon as possible. Although a few problems have already been made, such investigations have yet to be carried out. Even though the rules differ, NT limits must include military and civilian sectors and systems due to their fundamental nature and potential misuse. Thus, national and international laws must be firmly linked. Nanotechnology may revolutionise medicine, and research is ongoing. Researchers have implanted DNA nanobots into cockroaches to deliver medications. Nanobots were put into mice's stomach linings. Will brain nanobots become a reality like flying cars? 2) Military obligation. In the 21st century, military technology is on display. What technologies can revolutionise the military in this new era? Mi robots will lead a worldwide war near our battlefield. Russian military analyst Ivan Chekikov said the revolution would enhance operational models, principles, and war tactics (Maguire, 2007),(Spece et al., 2014).

Military nations are heavily investing in nanobots and dozens of nanorobot components. Studies predict military nanorobots in 2025. It is unknown when they will be completed and how the international political army will affect them.

How do nanorobots create or destroy enemies? Does a group want to weaken the army? Their main methods are: Start by using nanobots to boost war weapon power. Second, create nano components that block faces, noses, mouths, and eyes.

Third, study novel chemicals or organisms for artificial or hybrid insects. These deadly microorganisms infect the enemy and troops.

Fourth, the nanorobot can self-replicate or self-propagate in the square camp after penetrating the adversary.

Nanobots can play offence and defence, although their defence is far better. John Alexander, a famous American military expert, says nanomachines create the circumstances for the strategic offensive as they take on more combat duties, notably in strategic defense.

6. Future Work

After reading about military nanotechnology, we realised it would be used for good and bad. This articleshows that nanotechnology is a constantly evolving science, as demonstrated in the development new weaponry, particularly missiles and surveillance nanosatellites. Military nanotechnology will also improve police protection. Nanotechnology safety research grows annually. However, nanoparticles manufactured intentionally and their health and environmental implications have been the focus. Nanotechnology's environmental impact and social security and moral risks have not been sufficiently studied to secure its benefits.Society should value these human traits:

1) Use the mechanism to examine nanomaterials' health and environmental consequences as soon as possible. This nanomaterial has been tested for human and environmental health, even if most research doesn't explain the toxicological process.
2) Nanomaterial detection methods are examined. Nanoparticles in the environment are unknown, but this information can be utilised to set environmental quality standards and mitigation methods. Rapid nanomaterial concentration detection is crucial.
3) Improve cross-disciplinary collaboration and nanotechnology's social security and ethics hazards. Pakistan lacks nanotechnology standards.

Criminals who exploit system flaws and imperil social security threaten peace. Science and technology are unknown. Hence the risks and advantages are unanticipated. They also prioritise technology over social issues. Well-known and well-researched technology dangers have significant psychological, societal, and economic implications. Nanotechnology flaws have increased social risks.

Nanoscale gadgets may smarten agricultural systems. Nanotechnology helps and hurts agriculture. Nanotechnologies' health effects are unknowable. Before putting these applications into place, the long-term effects of farmers' exposure to nanomaterials, their unknown life cycles, their interactions with the living or nonliving environment, and their ability to build up in the body should be studied. Nanotechnology hasn't been commercialised because of high processing costs, problems with scaling up R&D for prototypes and industrial production, and the public's perception of environmental, health, and safety risks. Governments should establish tight laws and monitoring before commercialising and using nanomaterials. Governments must educate the important industry on nanotechnology's risks. Nanotechnology should be regulated shortly. Nanotechnology will spread. To avoid a massive industry revival, we must act. Nanorobots for future warfare require a new HIT-like industry.

7. Conclusion

We must be realistic about nanotechnology's progress. Remember the potential repercussions of rapid development when researching and creating nanotechnology. We can't cure grams in America.

As a result of the "millennium bug" hysteria, we must halt computer technology growth and prevent the global economy from losing 600 billion dollars. Nobody can be negative about nanotechnology, just like nobody can be negative about history.

This effect limits its complete development. However, remember that nanotechnology is still quite young. Before harnessing the synchronised expansion of nature and society, considerable steps must be taken to promote the general health of nanotechnology. Preventing misapplication and catastrophe at the source. Nanotechnology has changed technology, industry, and daily life, yet it has also had unanticipated effects. However, similar to genetic engineering, nanotechnology cannot be disregarded. Environmental, health, and social security concerns must be addressed to ensure nanotechnology's longevity. The government and the military should be concerned about the expansion of nanotechnology.

Governments must disseminate their expertise in the relevant disciplines to avoid or reduce the risks of nanotechnology. Establish appropriate nanotechnology regulations or limits as soon as possible. In the future, nanotechnology will certainly be utilised more frequently. Therefore, we must be willing to take measures to prevent returning to the old heavy industry path. We require a new sector similar to HIT Pakistan to produce nanorobots for next-generation warfare.

This article intends to increase global awareness of the hazards posed by future research on these concerns and the necessary preventative weapons control measures.

References

Ali, A., Qasim, M., Dilawar, M. U., Khan, Z. F., Jadoon, Y. K., & Faiz, T. (2022). Nanorobotics: Next level of military technology. *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1–7.

Ali, A., Hafeez, Y., Hussain, S., & Yang, S. (2020). Role of requirement prioritization technique to improve the quality of highly-configurable systems. *IEEE Access*, *8*, 27549–27573.

Thomas, S., Ahmadi, M., Nguyen, T. A., Afkhami, A., & Madrakian, T. (2021). *Micro-and Nanotechnology Enabled Applications for Portable Miniaturized Analytical Systems*. Elsevier.

Ali, A., Hafeez, Y., Hussainn, S. M., & Nazir, M. U. (2020). Bio-inspired communication: A review on solution of complex problems for highly configurable systems. 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET), 1–6.

Huang, C., Anderson, J., Peana, S., Chen, X., Ramanathan, S., & Weinstein, D. (2021). Perovskite Nickelate Actuators. Journal of Microelectromechanical Systems, 30(3), 488–493.

Caliskan, Z., & Sokullu, R. I. (2019). Drone Based Hotspot Network System Design. 2019 IEEE 30th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops), 1–6.

Stoudt, D. C. (2012). Naval directed-energy weapons-No longer a future weapon concept. NAVAL SURFACE WARFARE CENTER DAHLGREN DIV VA.

Jesudoss, A., Vybhavi, R., & Anusha, B. (2019). Design of smart helmet for accident avoidance. 2019 International Conference on Communication and Signal Processing (ICCSP), 0774–0778.

Ali, A. (2021). Artificial Intelligence Potential Trends in Military. Foundation University Journal of Engineering and Applied Sciences (HEC Recognized Y Category, ISSN 2706-7351), 2(1), 20–30.

You, Z. (Ed.). (2018). Chapter 2—Multidisciplinary Design Optimization of a Micro/Nano Satellite System. In Space Microsystems and Micro/nano Satellites (pp. 51–74). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-812672-1.00002-3

Ali, A., Said, R. A., Rizwan, H. M. A., Shehzad, K., & Naz, I. (2022). Application of Computational Intelligence and Machine Learning to Conventional Operational Research Methods. 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 1–6.

Debnath, M. (2016). Protection of inter-continental ballistic missile (ICBM) from anti-ballistic missile (ABM) by using anti-anti-Ballistic missile (AABM). International Journal Of Engineering And Computer Science.

Baber, Z. (2004). “An undifferentiated mass of gray goo?” Nanotechnology and society. Bulletin of Science, Technology & Society, 24(1), 10–12.

Goyal, A. K., Rath, G., & Garg, T. (2013). Nanotechnological approaches for genetic immunization. DNA and RNA Nanobiotechnologies in Medicine: Diagnosis and Treatment of Diseases, 67–120.

Kita, R., & Dobashi, T. (2015). Introduction of Nano/Micro science and technology in Biorheology. Nano/Micro Science and Technology in Biorheology: Principles, Methods, and Applications, 1–6.

Ali, A., Ghouri, K. F. K., Naseem, H., Soomro, T. R., Mansoor, W., & Momani, A. M. (2022). Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security. 2022 International Conference on Cyber Resilience (ICCR), 1–5.

Fujitsuna, M. (2019). The 1950 Model Year CarEA type Electric Vehicle DENSO-GOU’. IEEJ Transactions on Industry Applications, 139(6), 574–579.

Maguire, R. (2007). The use of weapons: Mass killing and the United Kingdom government’s nuclear weapons programme. Journal of Genocide Research, 9(3), 389–410.

Spece, R., Yokum, D., Okoro, A.-G., & Robertson, C. (2014). An empirical method for materiality: Would conflict of interest disclosures change patient decisions. Am. JL & Med., 40, 253.

Olejníček, A., Odehnal, J., Holcner, V., & Krč, M. (2018). Determinants of the Military Robotics Proliferation. Advances in Military Technology, 13(1), 71–86.

Foundation University Journal of Engineering and Applied Sciences
Vol. 3, Issue 1, 2022.
DOI: 10.33897/fujeas.v3i1.687

Research Article

Intrusion Detection in Cyber Space Using Machine Learning Based Algorithm

Muhammad Bashir, Muhammad Atique, Saif Ur Rehman, Muhammad Ibrahim Khalil

University Institute of Information Technology (UIIT), PMAS Arid Agriculture University
Rawalpindi, Pakistan

Corresponding Author: Saif Ur Rehman. Email**:** saif@uaar.edu.pk

Abstract:

Now a day, the fast growth of Internet access and the adoption of smart digital technology has resulted in new cybercrime strategies targeting regular people and businesses. The Web and social activities take precedence in most aspects of their lives, but also poses significant social risks. Static and dynamic analysis are inefficient in detecting unknown malware in standard threat detection approaches. Virus makers create new malware by modifying current malware using polymorphic and evasion tactics in order to fool. Furthermore, by utilizing selection of features techniques to identify more important features and minimizing amount of the data, these Machine Learning models' accuracy can be increased, resulting in fewer calculations. In the previous study traditional machine learning approaches were used to detect Malware. We employed Cuckoo sandbox, a malware detection and analysis system for detection and categorization, in this study we provide a Machine Learning based Intrusion analysis system to calculate exact and on spot Intrusion classification. We integrated feature extraction and component selection from the file, as well as selecting the much higher quality, resulting in exceptional accuracy and cheaper computing costs. For reliable identification and fine-grained categorization, we use a variety of machine learning algorithms. Our experimental results show that we achieved good, classified accuracy when compared to state-of-the-art approaches. We employed machine learning techniques such as K-Nearest Neighbor, Random Forest, Support Vector Machine, and Decision Tree. Using the Random Forest classifier on 108 features, we attained the greatest accuracy of 99.37 percent. We also discovered that Random Forest outscored all other classic machine learning techniques during the procedure. These findings can aid in the exact and accurate identification of Malware families.

Keywords: Cyber security; Security issues; Malware attacks; Cyber space; Intrusion detection

1. Introduction

The Internet space became an increasingly popular source of data as well as services. Internet usage rapidly increased after 2017, over 48 percent of the global population used the Internet as a channel of information. (Zhu, Jang-Jaccard, & Watters, 2020) In developed countries, this ratio rose to 81 percent.

The Internet's principal function is to carry data from one point to another across a network. The use of the Internet has risen dramatically as a result of advancements in computer systems, networks, and mobile devices. (Aslan & Samet, 2020) As a result, cybercriminals and adversaries have turned their attention to the Internet. Information confidentiality, availability, and integrity must all be guaranteed via a safe and stable computer system.

The term "cyber security" refers to a collection of security procedures that can be used to secure cyberspace and user assets against unwanted access and attacks. (Roseline, Sasisri, Geetha, & Balasubramanian, 2019) The basic goal of a cyber defense system is for data to be secure, integral, and accessible. Computer networks are (or should be) intended to provide safeguards that restrict data access to just those who are authorized. Threat makers have begun to make threat online rather than in the real world as the Internet has grown in popularity. (Shaukat, Luo, Varadharajan, Hameed, & Xu, 2020) Digital data, machines that handle digital information, and programmed that manipulate digital data are all protected by cyber security.

Malicious software can be classified in a variety of ways. Zhu et al. (Zhu et al., 2020) categorized malware into three categories based on how harmful software distributes, including worms, viruses, and trojan horses. Another classification is based on the behavior of malicious programs after it has successfully infected the victim's computer. Ravi et al. (Ravi Vinayakumar et al., 2019), Aslan et al. (Aslan & Samet, 2020), and Ulalh et al. (Ullah et al., 2019) gave a more thorough reference to kinds of harmful software, dividing malware into classes depending on its utility and the manner it affects machines.

- **Virus –** Viruses are a sort of malicious code that is not self-contained and is frequently attached to other programs (commonly.exe files). Viruses can propagate throughout a system or network, infecting other programs and machines, thanks to the replication feature. The virus is carried out by an unknowing user who want to run a program without first confirming its source. Pure computer viruses, according to Ravi et al. (Ravi Vinayakumar et al., 2019) are less widespread than in the past and now account up only 10% of all malware. Viruses are difficult to eradicate from the system due to their propagation role, prompting antivirus programs to delete infected files entirely (Roseline et al., 2019). Viruses deplete system resources and can result in a service interruption.
- **Worm –** Unlike viruses, worms may exist on their own and have self-replicating capabilities, allowing them to infect all machines in a network without the need for human involvement or an infected application. The worm spreads via email and storage devices. Because of these characteristics, worms are extremely dangerous, as a single employee who views an infected email or uses a lost memory stick might infect the entire corporate network (Gülmez & Sogukpinar, 2021). Stuxnet, one of the most well-known computer viruses, was used to disrupt Iran's nuclear program by attacking supervisory control and data acquisition (SCADA) systems.
- **Trojan –** Horse unlike earlier examples, the Trojan horse is incapable of self-propagation. It is transmitted through the internet as a download file or malicious attachments in emails and needs human contact to attack machines. Trojans are made to appear as if they are a real and innocuous software. "The most common Trojan type is a false antivirus program that appears and claims your computer is infected, then instructs you to run a program to clean it." (Ravi Vinayakumar et al., 2019).
- **Adware –** Adware is computer virus which does not seriously affect the user or the machine, but is instead used to earn revenue for the attacker by displaying advertisements. Some adware directs users to websites that contain harmful software and encourages them to download it. Adware is frequently included in free programs. Sehatbakhsh et al. (Sehatbakhsh et al., 2019) investigated whether adware should be classified as malware. According to the authors, it is difficult to quantify the security risks posed by adware, and the community is currently arguing whether or not adware should be classified as malware.
- **Spyware –** Spyware is a persistent sort of malware designed for stealthy, long-term operation on compromised machines, according to authors (Xu, Xia, & Shen, 2020). Spyware captures personal data, such as passwords and financial information, and sends it to the attacker without

the victim's knowledge or agreement. According to Sreekumari (Sreekumari, 2020), spyware is frequently utilized in the reconnaissance phase of more complex attacks. It's frequently linked to whaling, espionage, and data theft for blackmail (Davis, 2019). Table 1 shows the different notations used.

Table 1: Notations definition table

Notation	Description
ML	Machine Learning
DNN	Dynamic Neural Network
NSL	Network Security Laboratory
UNSW-NB15	Network Intrusion Detection System
WSN-DS	Wireless Security Network- Data Set
NIDS	Network Inventory and Design System
HIDS	Host Based Intrusion Detection

We present a malware analysis approach in this paper to find and classify malware more accurately. We implemented the feature extraction module to extract features from the analysis report for malware. Furthermore, most important characteristics are extracted from the collected features and used to construct training and testing datasets using multiple feature selection techniques. Finally, we run the dataset using various Machine Learning (ML) methods. We employed Cuckoo sandbox, an automated malware analysis system for detection and categorization, in this study to provide a Machine Learning based Intrusion analysis system to calculate exact and on point Intrusion identification. We integrated feature extraction and component selection from the file, as well as selecting the much higher quality, resulting in exceptional accuracy and cheaper computing costs.

The next sections of the paper describe the previous related research work as literature review and then the proposed methodology of the framework we used in our research. The next section presents the experimental result and the finding of our research. After this, we discussed the outcomes and concluded our research with our discussed findings.

2. Literature Review

This section gives a comprehensive summary of recent work on machine learning in cyber security. We additionally narrowed our search by considering algorithm details and efficiency, feature extractions and selection methods (if applicable), and the relevant dataset used to solve an issue (s). This document also includes a quick summary of all the strategies.

A Multi-Loss Siamese Neural Network with Batch Normalization Layer was suggested by (Zhu et al., 2020), which can identify more accurately with fewer samples. Their model, which is trained with only a few samples, uses the Siamese Neural Network to detect new malware types. Their model includes batch normalization and several loss functions to handle over fitting caused by tiny samples, which can result in a vanishing gradient problem due to binary cross-entropy loss, as well as embedding space to improve detection accuracy. In addition, they show how to transform raw binary data into malware grey scale images, as well as how to generate positive and negative pairs for training the popular Siamese Neural Network. Their results suggest that their model outperforms similar methods currently in use.

They presented a method for converting a binary file, such as the intrusion dataset, into raw images that may be fed into a neutrality network model, such as a Siamese Network. Their suggested model has been adjusted to operate well with a small number of training datasets. This is accomplished by appropriately adjusting the network parameters for feature extraction and applying batching normalization to prevent overfitting induced by the usage of tiny datasets. In the Siamese Neural Network for binary classification, the repeated error function is important for improving the feature embedding space. Each positive pair belonging to the same class has a tiny distance in the characteristic embedding. For one classification in malware detection, their suggested model outperforms the standard methods, according to their experimental data.

In another study, a variety of publicly available benchmark malware datasets, a complete finding of experiments with Dynamic Neural Network and other conventional machine learning classifiers (Ravi Vinayakumar et al., 2019). With the KDDCup 99 dataset, the ideal network Structure for Dynamic Neural Network are determined using hyper parameter selection approaches. All DNN trials are done for 1,000 epochs with learning rates ranging from [0.01-0.5]. To conduct the benchmark, the Dynamic Neural Network framework that over powered on KDDCup 99 is performed to various data files such as NSL-KDD, UNSW-NB15, Kyoto, WSN-DS, and CICIDS 2017. By feeding IDS data through several hidden layers, their Dynamic Neural Network model learns the multi-dimensional representation of feature. It has been proven through thorough experimental testing that DNNs outperform traditional machine learning classifiers. Finally, they presented Scale-Hybrid-IDS-AlertNet (SHIA), hybrid Dynamic Neural Network framework that will be utilized in run time to successfully discover network traffic and events in order to notify incoming cyber-threats. To avoid detection by the IDS, an attacker pretends to be a regular user. Intrusive behavior patterns, on the other hand, differ in some ways. This is related to an attacker's specific goal, such as gaining illegal entry to hardware and communications resources. The trend of internet resource usage can be captured, but current approaches have a significant false positive rate. Intrusion patterns can be found in typical traffic with a low key across long spans. An effective deep learning strategy is provided by building a deep neural network (DNN) to identify cyberattacks proactively by merging both NIDS and also HIDS jointly. On various NIDS and HIDS datasets, the efficacy of various traditional machine learning techniques and DNNs in identifying whether internet traffic behavior is normal or anomalous related to an attempt which can be categorized into appropriate attack categories is investigated in their paper.

$$h_i(x) = f(w_i^T x + b_i)$$

The mathematical formulae for sigmoid and tangent are given below.

$$Sigmoid = \frac{1}{1+e^{-x}}$$

$$Tangent = \frac{e^{2x}-1}{e^{2x}+1}$$

Authors (Shaukat et al., 2020) proposed that their paper aims to provide a comprehensive overview of the challenges that ML techniques face in protecting cyberspace against attacks by presenting a literature on ML techniques for cyber security, including threat detection, spam filtering, and intrusion detection on computer networks and mobile networks over the last decade. It also includes brief definitions of each machine learning method, as well as security datasets that are often utilized, fundamental machine learning tools, and assessment metrics for assessing a categorization model. Finally, it analyses the difficulties of employing machine learning techniques in cyber security. This presentation includes a wide bibliography as well as recent ML trends in cyber security. ML techniques are used between sides of the attack, i.e., the attacker and the cyber security side. On the cybercriminal side, ML approaches are being used by malicious hackers and fraudsters to uncover system flaws and advanced attack methods to get past the defense wall. On the defense side, machine learning models are helping to give more robust and intelligent strategies for improving efficiency and early intervention of attacks, reducing the impact and damage.

Aslan and Samet (Aslan & Samet, 2020) presented that the Detection approaches based on signatures and heuristics are both quick and effective for detecting well-known threats also signature based approaches are not best to discover novel intrusion. For unknown and difficult intrusion, behavior-dependent, Neural learning-based, smart phone, and Internet - of - things technologies are also emerging to detect some portion of predictable and unpredictable threats. Model-checking-based and cloud-based systems perform well. However, no approach can detect all malware in the wild. This illustrates that finding an effective method for detecting malware is a difficult endeavor, and that new research and methodologies are desperately needed. The proposed study examines malware detection technologies in depth, as well as recent detection methods that use these approaches. The purpose of this methodology is to provide researchers with a broad understanding of malware detection approaches, as well as the benefits and drawbacks of each detection methodology and the methods employed in these approaches. Malware must be studied in order to comprehend its content and actions.

Roseline et al. (Roseline et al., 2019) suggested that in comparison to deep learning models, a hybrid stacked multilayered ensembling technique is more resilient and economical. With an accuracy of 98.91 percent, the suggested approach beats machine learning and deep learning models. The recommended technique works well for both little and large-scale data due to its versatility in automatically altering parameters (number of consecutive levels). In terms of resources and time, it is computationally efficient. When compared to deep neural networks, the approach requires far less hyper-parameters. Pattern analysis through visualization is used as an alternate approach of analyzing malware. Malware samples are transformed to grey images, which allow different classes to see distinct patterns. This method does not necessitate the execution of samples. They don't necessitate any special skills or monitoring software.

Farhan et al. (Ullah et al., 2019) proposed that, to detect source code plagiarism, a deep learning algorithm is deployed. The data was gathered as part of the Google Code Jam (GCJ) investigation into software piracy. Aside from that, through color image visualization, the deep convolutional neural network is used to detect dangerous infections in IoT networks. Malware samples were gathered from the Maling dataset for testing purposes. The experimental results show that the suggested solution's classification performance for measuring cybersecurity threats in IoT is superior to current methodologies. Traditional approaches may address code concealment issues, however texture feature mining with virus visualizations requires a significant computational cost. With substantial malware data analysis, these kinds of data mining algorithms do not perform effectively. Virus is presently constantly producing, updating, and manipulating itself, making identification more difficult. The suggested malware identification system aims to answer the questions below: How can malware be identified with minimal effort? How can malware traits be extracted with a lower computational cost? How to improve accuracy by processing large malware datasets.

Sethi et al. (Sethi, Kumar, Sethi, Bera, & Patra, 2019) presented that they created a ML based intrusion analysis system for improved and precise intrusion identification and categorization. They used framework for multi-dimensional identification, which runs intrusions over a separate space also creates a report according to behaviors. They also performed extraction of features with implementing module that retrieve features from file and selects the most special characteristics to ensure more improved accuracy while lowering power costs. They utilized a number of ML methods for precise identification and streamed categorization. By this they achieved good identification and also categorization precisions when compared to high end approaches.

Gulmez et al. (Gülmez & Sogukpinar, 2021) proposed that the prevention detection methods based on static analysis, most malware makers employ obfuscation and encryption techniques. This type of malware is known as packed malware, and it is widely believed that in order to identify it, it must either be unpacked or evaluated dynamically. To address these issues, a graph-based malware detection method is suggested. The proposed method works by getting the opcode graph of each executable file in the dataset and using it to extract data later. As a result, the proposed approach achieves a detection accuracy of up to 98 percent. Aside from the high accuracy rate, the suggested method allows for the detection of packed malware without the requirement for unpacking or dynamic analysis. Graphs are used to represent source code. The classification of the graphs, they believe, leads to the discovery of malware. Different methods have been implemented to achieve this goal. The initial step is to deconstruct all of the dataset's files. The opcodes of a file are exposed during deconstructing, and these opcodes create opcode series.

Zhang et al. (Zhang, Kodituwakku, Hines, & Coble, 2019) suggested that Traditional process monitoring systems look for cyber-threats that are not identifiable by supervising network, such as data manipulation and false input attacks by an inside user. Regression model is investigated in the suggested detection system to increase early assault detection. The results suggest that following method identify damaging cyber threats as soon as they have a significant impact. A viable method for protecting an ICS is the suggested multiple layered input driven intrusion identification system, which uses data as well as networks and system. The proposed cyber-attack detection system's architecture, based on the defense and security concept. The classic intrusion detection and prevention layer, which includes firewalls, data diodes, and gateways, is the initial protection layer and is already widely used in the industry. However, in some cases, the attackers may be able to get beyond this defense line. The second security layer is made up of information models for detecting cyber-attacks based on network traffic and system data, such as the M1 classification algorithm and M2 big data models.

Vinayakumar et al. (Vinayakumar, Alazab, Soman, Poornachandran, & Venkatraman, 2019) presented that the first part of their paper compares and contrasts traditional MLAs as well as deep learning framework for intrusion identification and also categorization utilizing available and restricted datasets. By employing distinct portions of the available and restricted datasets to learn and run the model in a separable manner using timelines, they eliminated all data biasness from the experimental results. They proposed a unique visual processing technique that uses best settings for MLAs as well as deep learning framework to produce a precise intrusion identification model. They suggested deep learning architectures outperform classical MLAs, according to a comprehensive comparison assessment of our model. Their innovation in integrating visualization and deep learning architectures for a hybrid method based on static, dynamic, and image processing performed in a big data platform is unique in the world in terms of providing robust intelligence zero-day threat detection.

The author of (Sehatbakhsh et al., 2019) suggested that they offer REMOTE, a new structure that is meant to address logistical problems for tracking resource-constrained devices (e.g., embedded devices, IoTs, CPS, and so on), such as: Availability of code, assessment, and/or instruments equipment: source code, measurement, and/or instrumentation infrastructure may be unavailable. Flexibility in Software as well as Hardware: the equipment may be determined by different CPU architectures, and the device may use different operating systems or just not be there at all, Containers may limit immediate access to the network, such as placing a power or EM sensor very near to a microprocessor or indeed the main panel. When employing analogue signals for surveillance, the environment can modify the signal that is sent out and/or add interference to the signal that is received.

Xu et al. (Xu et al., 2020) proposed that When it comes to online attacks, earlier research has all assumed that virus will carry out attacks on the electric grid as soon as it infects a new host. In fact, the virus will not initiate attacks until enough hosts have been compromised in order to achieve the best attack effect. The virus has a lag phase during which it causes no damage but merely spreads quietly to infiltrate victims and evade detection (for example, in the Ukrainian incident, the virus had already been spreading surreptitiously in the telecommunications for more than six months before carrying out attacks). Previous research has rarely analyzed in depth the implantation duration and malware detection likelihood, both of which are important elements in determining the level of harm caused by malware attacks in CPPSs. As a result, we suggest a more practical paradigm in this brief that takes these overlooked elements into consideration.

Javaheri et al. (Javaheri, Lalbakhsh, & Hosseinzadeh, 2021) presented that the method to identifying unusual virus types, as well as present and prospective alterations produced by tectonic algorithms Several of the malware classes their studied are highly rare and cryptic. From the Adminus, VirusSign, and VirusShare computer viruses' datasets, they were able to collect and acquire some important data for these classes. They selected examples from different types of unusual malware at randomly to construct the first population, comprising stealthy spyware, kernel rootkit, injector, blocker, bootkit, evader, and file-less malware. To evade suspicion and exposure of their reaction, all samples from the early population malware were highly packed and secured by tough packers, as well as various unknown packers. The initial population is insufficiently large to allow for good and reliable categorization. This difficulty was solved by creating a fresh dataset, which increased the population's size and quality, allowing for more reliable training.

Sreekumari et al. (Sreekumari, 2020) suggested that By collecting feedback' device action without one‘s approval, types of malware can play a number of roles, including theft, encoding, erasing, and changing personal details such as bank account information, credit card information, login information, Security Numbers, and so on. Furthermore, malware can alter the system's needed functions by adding, changing, or removing programs. It takes advantage of a digital system's weaknesses, which are flaws or risks of harm. Malware's delivery reveals the malware's intent. Hackers write payloads, which are little pieces of code that operate on the machines they infect. Whereas many people are delighted with their modern technological presents and technology, they may be unaware of the frightening rate at which attackers are hacking their gadgets. Deep learning, for example, is one of the most essential ways for identifying malware because it has been shown to have a positive impact on natural language processing and image categorization.

Further, the comparison of closely related techniques is presented in Table 2.

Table 2: Comparative analysis of malware detection technique

Article	Methodology	Research Findings	Dataset Used	Malware Addressed	Accuracy
Jang-Jaccard et al. (Zhu et al., 2020)	Multi loss Siamese neural network	Batch Normalization and multi loss function	VirusTotal Dataset	All malware family	99.2%
Vinayakumar et al. (Ravi Vinayakumar et al., 2019)	Deep neural network	Scale-Hybrid-IDS-Alert Net	HIDS, NIDS KDDCup 99	All malware family	92.1%
Shaukat, K et al. (Shaukat et al., 2020)	Machine learning Techniques	Applications of ML Models	DARPA ID	All malware family	94.1%
Aslan, O et al. (Aslan & Samet, 2020)	Malware detection approaches	Signature and heuristic based approaches	NSL-KDD, Drebin, EMBER	All malware family	97.1%
Roseline et al. (Roseline et al., 2019)	ML Random Forest	Effective with contemporary deep learning model.	Kaggle's BIG 2015	All malware family	98.8%
Ullah, F et al. (Ullah et al., 2019)	Deep learning Convolutional neural network	Deep learning for identification of pirated and malware files	Google code jam, Malimg	All malware family	96%
Sethi, K et al. (Sethi et al., 2019)	Machine learning Algorithm	Decision tree out performed other ML Algorithms	VirusTotal, VirusShare	All malware family	99.1%
Gulmez et al. (Gülmez & Sogukpinar, 2021)	Graph based operational codes	Proposed model out performed other techniques	VX-Heaven	All malware family	97%
Zhang, F. et al. (Zhang et al., 2019)	Machine learning Algorithm	High detection accuracy and wide attack coverage.	MITM	All malware family	98.8%
Vinayakumar et al. (R Vinayakumar et al., 2019)	Deep learning techniques MLAs	Deep learning out performed classical MLAs	Malimg	All malware family	98.9%
Sehatbakhsh et al. (Sehatbakhsh et al., 2019)	REMOTE	REMOTE out performed state of the art external malware detection framework.	MiBench	All malware family	99.8%
Xu, S et al. (Xu et al., 2020)	Cyber Physical Power System (CPPS)	Defer the maximum payoff of the attacker.	CPPS	----	97.8%
Javaheri, D et al. (Javaheri et al., 2021)	Genetic Algorithm	Genetic algorithm out performed.	Microsoft, Malimg	All malware family	98%
Sreekumari et al. (Sreekumari, 2020)	Deep learning Algorithms	Deep learning is the profound solution for Malware detection.	Maling	----	96-99%

3. Proposed Methodology

In this portion, we present a full overview of our suggested framework for identifying and categorizing provided data.

We created a Python script to extract significant features, which was then used to choose the most important features, resulting in training as well as testing data pool. The data pool consists of a Mega data pool used for identification as well as a Mini data pool for categorization. For the identification and categorization of the given dataset, we employ multiple machine learning methods offered by the Scikit-learn module in Python. Scikit-Learn is a popular machine learning library in Python that provides a range of supervised and unsupervised learning algorithms. The library includes models based on machine learning approaches such as linear models, tree-based models, support vector machines, naive Bayes, clustering, and neural networks. Scikit-Learn also provides various tools for data preprocessing, model selection, and evaluation. The proposed methodology's operational flow is depicted in Figure 1. There are three steps to it as listed in Algorithm 1. We provide a full description of these three stages in this article.

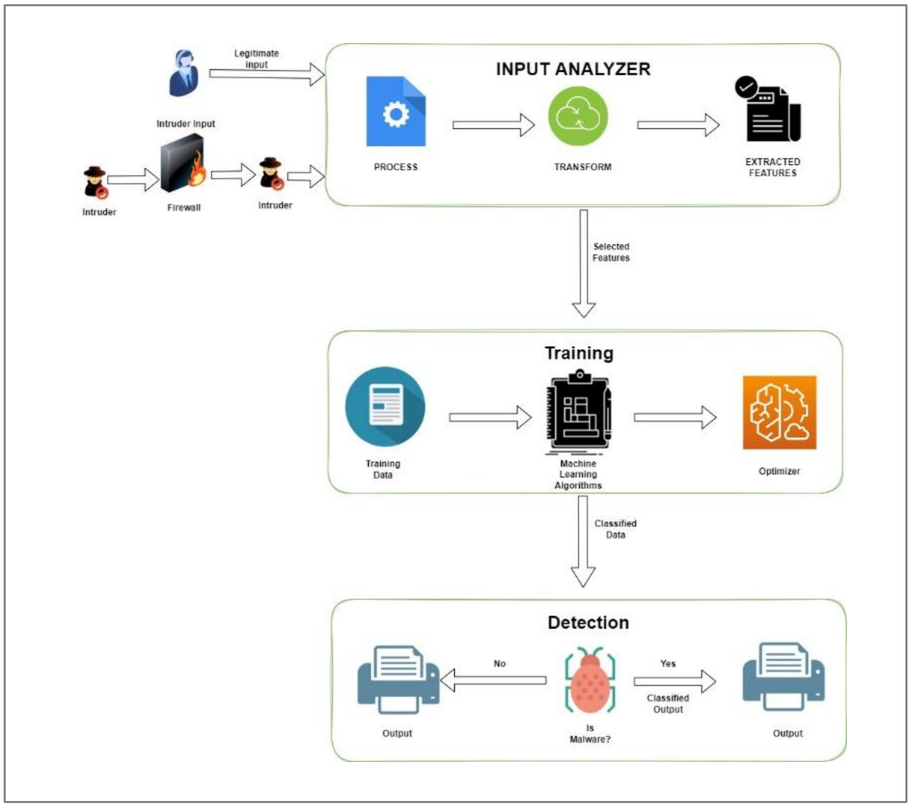


Figure 1: Structural design of implemented framework

3.1 Input Analyzer (Data Collection)

Due to the lack of a comprehensive dataset for malware analysis, we created our own. We'll use this dataset to conduct more malware analysis utilizing machine learning algorithms. We gathered malware and clean files containing windows PE header files from the VirusShare and VirusTotal websites.

3.2 Training (Pre-processing and Feature Extraction)

Because each source code has a different syntax and semantic structure, detecting pirated software among several types of source codes is a difficult operation. Detecting pirated software can be challenging

***Algorithm** 1 Proposed Methodology Steps*

```
Let ζD= dataset
Begin
Step 1: Get(ζD)
Step 2: Input Analyzer
Step 3: Training dataset
   3.1. Dataset splitting for training, testing and validation
   3.2. Feature Extraction EfficientNetB0 pre-trained model
   3.3. Optimize (epochs, batch size learning weights)
Step 4: Evaluation Metrics (accuracy, precision, F1 score and recall)
End
```

because pirates use various techniques to hide their activities and disguise the origin of the software. Some common techniques include code obfuscation, modifications, encryption, distribution methods, and changing file names. These tactics make it difficult to identify pirated software among legitimate sources. However, software companies and law enforcement agencies are continuously developing new methods to detect and prevent software piracy. To detect source code similarity, we used software plagiarism approaches. The source codes are broken down into little parts for deep analysis using pre-processing techniques. We used Scikit Learn library to complete the Data Preprocessing. It covers stemming, root word extraction, and frequency extraction, as well as the elimination of words. It translates the codes into relevant information and filters out the noise. Unwanted features, such as special symbols, constants, and stop words, are removed from the data. The cleansed data is then transformed into useable tokens using the tokenization process. In the pre-processing step, stemming, root words, and frequency limitations are employed to extract more valuable characteristics. The contribution of each token is then zoomed using weighting techniques. Data cleansing and transformation are important steps in data analysis because they help to ensure that the data being used is accurate, complete, and relevant to the research question at hand. When working with data, it is common to encounter missing values, outliers, and other types of errors or inconsistencies. These issues can affect the results of statistical analysis and make it difficult to draw meaningful conclusions from the data.

Weighting techniques are often used to adjust the importance of different observations in the data based on various factors such as sample size, response rate, or demographic characteristics. However, if the data is not cleansed and transformed prior to applying weighting techniques, the resulting weights may be biased or inaccurate. For example, if there are missing values or outliers in the data, these may affect the weighting calculations and lead to incorrect results.

Additionally, data cleansing and transformation can help to ensure that the variables being used are in the appropriate format and have the desired level of granularity. For example, if a variable is recorded in a different unit of measurement than what is needed for analysis, it may need to be transformed before applying weighting techniques.

In summary, data cleansing and transformation are important steps in data analysis that can help to ensure the accuracy and relevance of the data being used. By preparing the data properly before applying weighting techniques, researchers can increase the reliability of their results and draw more meaningful conclusions from the data.

In the weighting step, the TFIDF and Logarithm of Term Frequency (LogTF) are utilized. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical weighting scheme commonly used in information retrieval and text mining. It is used to represent the importance of each word or term in a document based on its frequency of occurrence across a collection of documents.

The term frequency (TF) component of TF-IDF measures the frequency of a word or term within a single document. The inverse document frequency (IDF) component measures how rare or common a word is across the entire collection of documents. The idea behind IDF is that words that occur frequently in many documents, such as "the" or "and", are less important in distinguishing between documents and should be given less weight, while words that occur rarely or only in a few documents are more important in distinguishing between documents and should be given more weight. LogTF, on the other hand, is a modification of the TF component of TF-IDF. It takes the logarithm of the raw frequency of a term in a document, which helps to reduce the impact of very high term frequencies. This is useful because when a term occurs very frequently in a document, it can skew the weight of that term and make it seem more important than it really is.

Stemming and root word extraction are techniques used in natural language processing to transform words into their base or root form, which can help to reduce vocabulary size and improve text analysis. Stemming involves removing suffixes to get to the base form of a word, while root word extraction identifies the base form without necessarily removing all suffixes. Both techniques can improve text analysis, but their effectiveness may vary depending on the language, domain, and task at hand.

3.3 Detection and Classification

The Scikit-Learn library was utilized to identify and categorize malware in our Implemented methodology. Marco dataset and Micro dataset are two datasets created utilizing feature purification. These datasets are then separated into learning and implementing sub datasets in a 70/30 parts for model learning and implementation, respectively. Used a python library of Scikit, a Model was created using machine learning approaches, the training data pools were called into a coded program of Python language. The models created in Scikit library are: (1) -Binary Classification: - A set of data is used to identify whether a provided sample is malicious or not; and (2)- Multi-Class Classifier: - Its extracted features or set of data is utilized to categorize a provided sample data into various virus kinds.

4. Experiment and Results

In this section, with detailed experimental findings, we'll show you how well our malware analysis process performed. For our investigation, we used 1200 data samples, 678 harmful samples and 5+ harmless samples were found. For both binary and multi-class classifiers, the Scikit-Learn module gave results with precise identification of each class and a matrix. We separated them into two groups for malware analysis: training and testing. Seventy percent of malware samples are in the training set, while thirty percent are in the testing set. Using the models created from the machine learning techniques, we observed results for identification and categorization. The following subsections explain the outcomes of the experimental examination.

4.1 Result Analysis

To analyze the effectiveness of our proposed malware analysis system, Reliability, Precise, Retention, F-measure, and Region underneath Fitted Model were the five performance measures we looked at. Real Positive, Untrue Positive, Real Negative, and Fake Negative are the performance measurements.

4.2 Evaluation Metrics

Before explaining the evaluation metrics, we would give a brief introduction to the confusion matrix. A confusion matrix is generally a 2x2 matrix layout that is beneficial for visualizing the performance of an algorithm. Actual performance measures that must or should be met are written vertically while the predicted ones by the algorithm are written horizontally (Ravi Vinayakumar et al., 2019).

True Positive Rate (TPR) is the total positive instances being identified as positive.

$$\text{TPR}=\frac{TP}{TP+FN}$$

True Negative Rate (TNR) is the number of negative instances being identified as negative.

$$\text{TNR} = \frac{TN}{TN+FP}$$

False Positive Rate (FPR) is the number of negative instances being classified or predicted as positive.

$$\text{FPR} = \frac{FP}{FP+TN}$$

False Negative Rate (FNR) is the number of positive instances being classified or predicted as negative.

$$\text{FNR} = \frac{FN}{FN+TP}$$

Accuracy: is the ratio between the number of correct predictions and a total number of predictions.

$$\text{Accuracy} =\frac{TP + TN}{TP+FP+FN+TN}$$

Precision: is the ratio between TPs combined to a number of TPs and FPs. It is the percentage of correctly identified positives out of all results which were said to be positive either correctly or not.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: is defined as the ratio between TPs combined to a number of TPs and FNs. It is the percentage of correctly identified positives out of all actual positives, either correctly or not.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-score: It takes both false negatives and false positives into consideration, and it is the harmonic mean of recall and precision. It performs well on datasets that are imbalanced.

$$\text{F1-Score} = 2 * \frac{(\text{precision}*\text{recall})}{(\text{precison}+\text{recall})}$$

5. Malware Identification Outcomes

Results of intrusion identification using several classifiers in the trial. Table 3 and Figure 2 shows that with 108 selected features, Random Forest has a high identification rate of 99.37 percent and an accuracy of 99.37 percent. Among other classifiers, this one has the highest accuracy. Similarly, we used Random Forest to attain high Precision, Recall, and F-Measure values. In comparison to the other classifiers employed in the experiment, these measures are the best. The comparison metrics are shown in Figure 3.

Table 3: Malware identification results

Algo	**Features**	**Accuracy**	**Precision**	**Recall**	**F-measure**	**AUC**
KNN	108	94.68%	0.95	0.95	0.91	0.85
Random Forest	108	99.37%	0.99	0.99	0.99	0.98
SVM	1000	89.37%	0.92	0.90	0.90	0.93
Decision Tree	92	95.93%	0.96	0.96	0.96	0.81

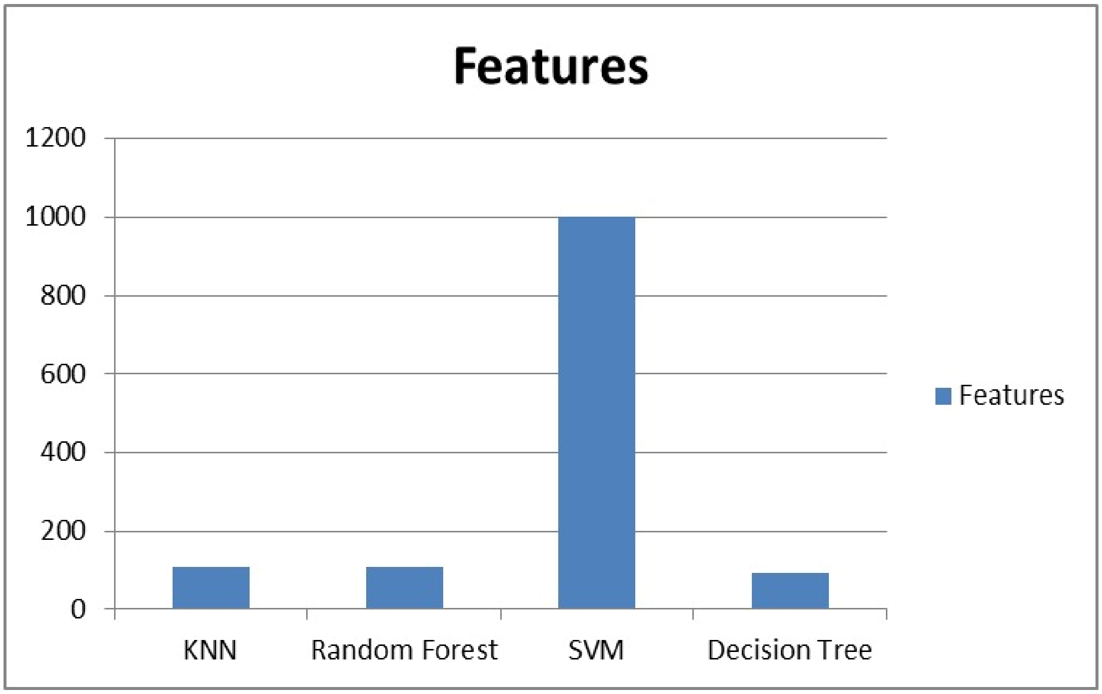


Figure 2: Graphical representation of features

Here next, different algorithms used for classification are briefly discussed.

5.1 K-nearest Neighbors

The function is only approximated locally in k-NN classification, and all computation is postponed until

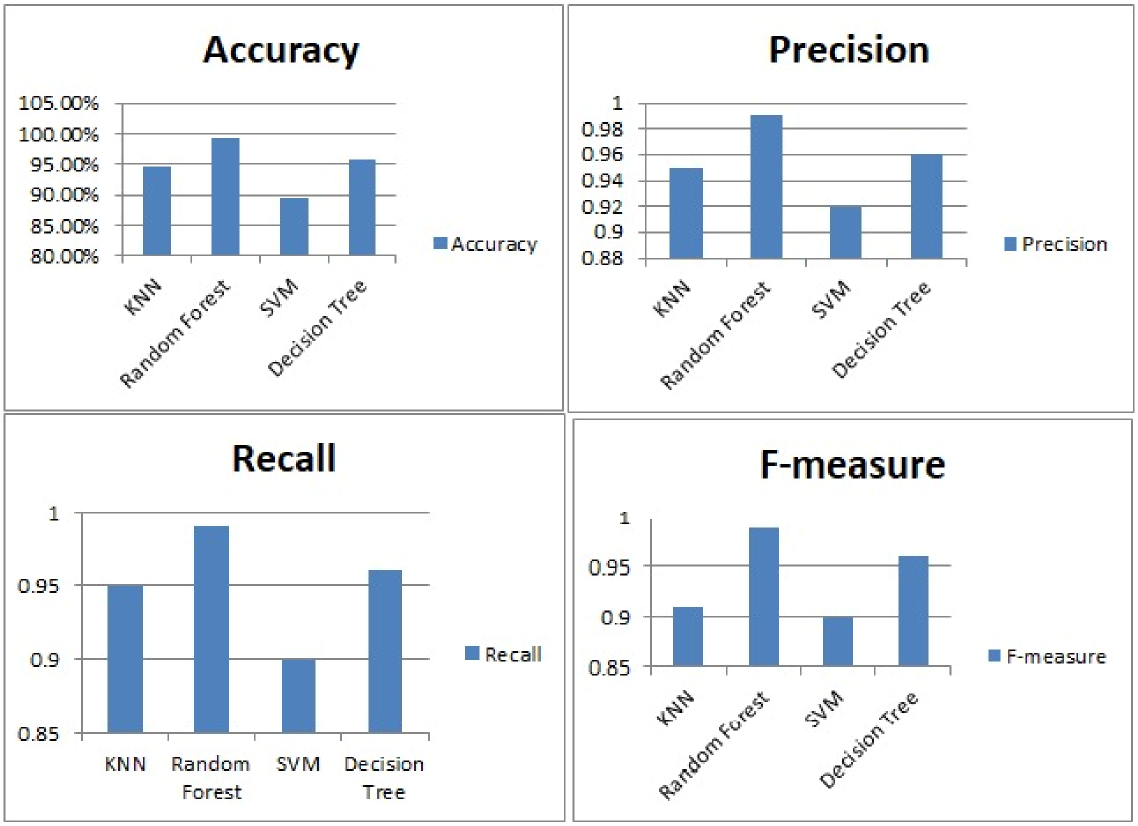


Figure 3: Graphical representation of experimental results

the function is evaluated. Because this method relies on distance for classification, normalizing the training data can greatly increase its performance if the features represent various physical units or come in wildly different scales. An effective strategy for both classification and regression is to assign weights to the contributions of the neighbors, so that the closer neighbors contribute more to the average than the farther neighbors. A popular weighting technique, for example, is to give each neighbor a weight of 1/d, where d is the distance between them.

In k-NN classification, the contributions of neighbors can be weighted to improve accuracy. One common method is the inverse distance weighting (IDW) method, which assigns weights to neighbors based on their distance to the point being classified. Another method is the Gaussian kernel weighting method, which uses a Gaussian function to weight the contributions of neighbors based on their distance. Both methods can be useful in cases where some neighbors are more informative than others due to their proximity to the point being classified.

5.2 Random Forest Algorithm

Random forests, also known as random choice forests, are an ensemble learning method for classification, regression, and other tasks that works by building a large number of decision trees during training. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. Random decision forests address the problem of decision trees overfitting their training set. Random forests outperform decision trees in most cases, but they are less accurate than gradient enhanced trees. Data features, on the other hand, can have an impact on their performance.

5.3 Support Vector Machine

Support-vector machines (SVMs, also known as support-vector networks) are supervised learning models that examine data for classification and regression analysis using related learning techniques. Vladimir Vapnik and colleagues developed it at AT&T Bell Laboratories (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995, Vapnik et al., 1997). SVMs, which are based on statistical learning frameworks or VC theory established by Vapnik (1982, 1995) and Chervonenkis, are one of the most reliable prediction approaches (1974).

5.4 Decision Tree

One of the predictive modelling methodologies used in statistics, data mining, and machine learning is decision tree learning, also known as induction of decision trees. It goes from observations about an item (represented in the branches) to inferences about the item's goal value using a decision tree (as a predictive model) (represented in the leaves). Classification trees are tree models in which the goal variable can take a discrete set of values; in these tree structures, leaves indicate classifiers and branching represent feature combinations that lead to those class labels. Logistic regression are decision trees in which the target variable can take continuous values (usually real numbers). Because of their comprehensibility and simplicity, decision trees are one of the most popular machine learning methods.

6. Malware Classification Results

6.1 Malware Categorization Results

The performance of classification utilizing the various classifiers employed in our experiment demonstrate that Random Forest has an extremely accurate percentage of 99.11 % in categorization. KNN, SVM, and DT, on the other hand, yield classification rates of 86.72 percent, 86.72 percent, and 88.23 percent, respectively. Random Forest also scored well in Precision, Recall, and F-Measure.

6.2 Malware Family Classification Results

Table 4 shows the malware family classification results for each class. have employed seven types of

Table 4: Intrusion family categorization results

Class	Samples	Accuracy	Precision	Recall	F-measure
Virus	81	90%	1.00	0.90	0.95
Adware	231	100%	1.00	1.00	1.00
Trojan	255	100%	0.94	1.00	0.97
Spyware	48	77.77%	0.58	0.78	0.67
Worm	27	66.66%	1.00	0.67	0.80
Backdoor	24	62.5%	1.00	0.62	0.77

malwares (viruses, adware, trojans, etc.) Backdoors, hoaxes, spyware, and worms are all examples of malicious software. Each class has its own set of rules. Figure 4 shows the number of samples we collected. For Adware and Trojans, the Random Forest classifier model worked best. Thus, Adware and Trojans viruses achieved the highest accuracy. Because of the enormous number of samples, malware types have been created. In comparison to other classes, for those classes. By achieving a high recall and precision rate for certain classes. This indicates that the projected results are very accurate. The results of evaluation metrics are shown in Figure 5.

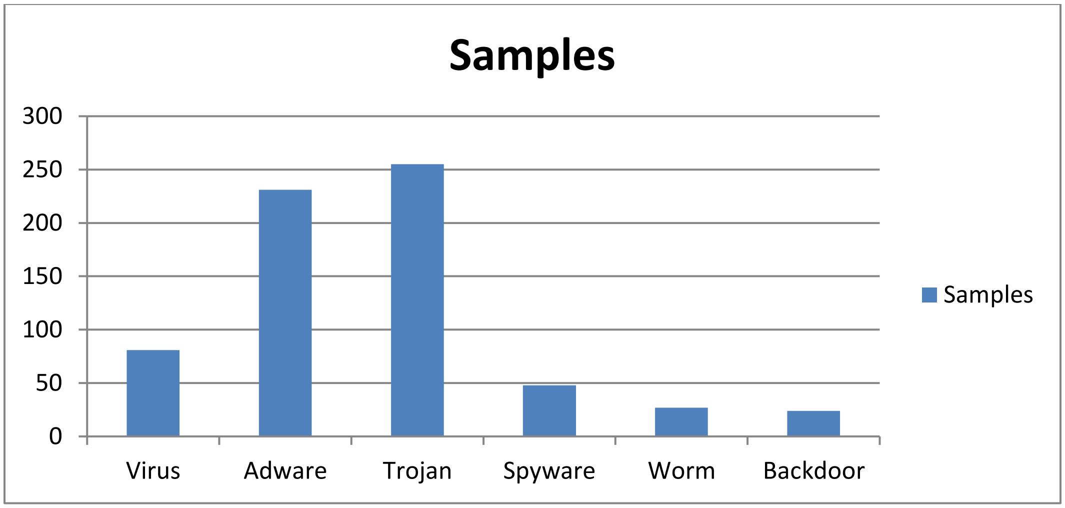


Figure 4: Graphical representation of samples

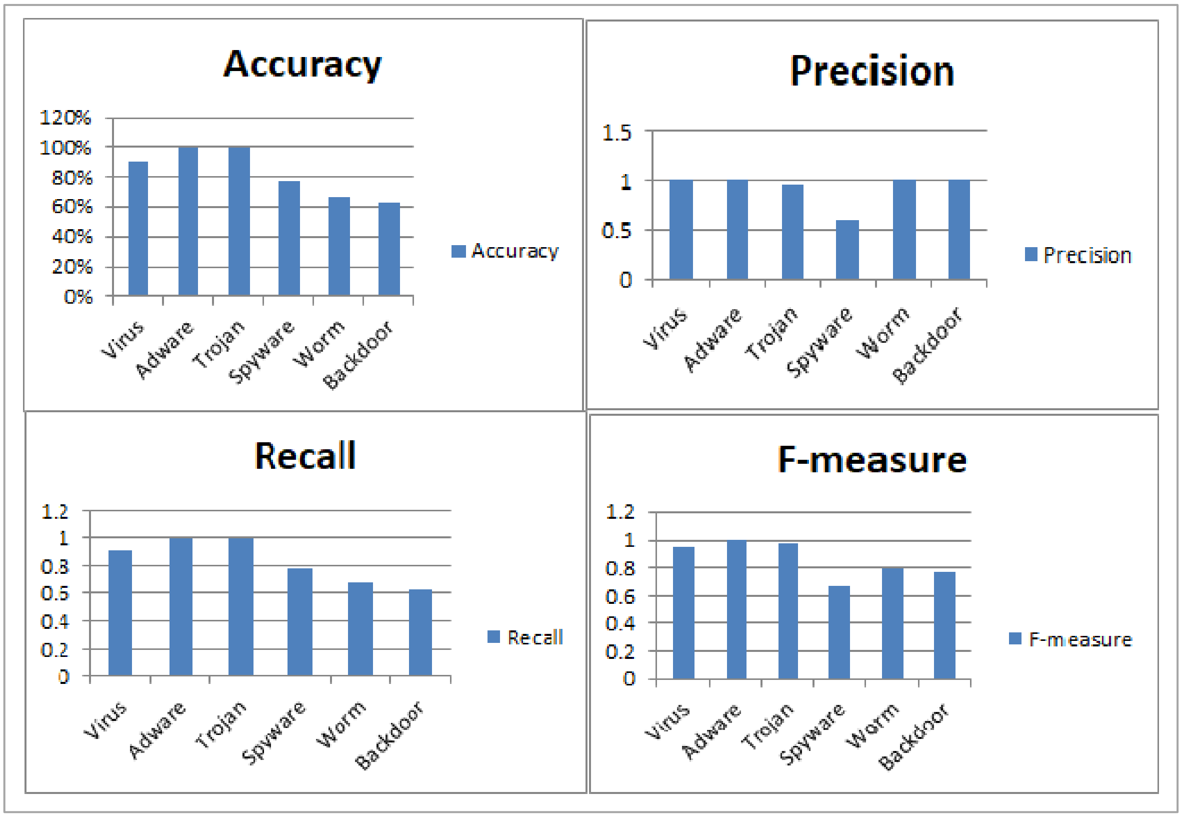


Figure 5: Graphical representation of results

7. Conclusion

We introduce a novel malware analysis methodology in this paper that can efficiently detect and categories malware. In order to extract the most important characteristics, our suggested approach uses two independent feature selection algorithms, this reduces learning time and enhances the reliability of detection and categorization. Gathered results reveal to with information like Random Forest has good detection and classification accuracy. Furthermore, we have classified intrusion based on its class and verified the identification of each intrusion class.

References

Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. *IEEE access, 8*, 6249-6271.

Davis, R. S. a. G. (2019). McAfee Mobile Threat Report Q1. from https://www.mcafee.com/enterprise/en-us/assets/reports/rp-mobile-threat-report-2019.pdf

Gülmez, S., & Sogukpinar, I. (2021). *Graph-based malware detection using opcode sequences.* Paper presented at the 2021 9th International Symposium on Digital Forensics and Security (ISDFS).

Javaheri, D., Lalbakhsh, P., & Hosseinzadeh, M. (2021). A novel method for detecting future generations of targeted and metamorphic malware based on genetic algorithm. *IEEE access, 9*, 69951-69970.

Roseline, S. A., Sasisri, A., Geetha, S., & Balasubramanian, C. (2019). *Towards efficient malware detection and classification using multilayered random forest ensemble technique.* Paper presented at the 2019 International Carnahan Conference on Security Technology (ICCST).

Sehatbakhsh, N., Nazari, A., Alam, M., Werner, F., Zhu, Y., Zajic, A., & Prvulovic, M. (2019). REMOTE: Robust external malware detection framework by using electromagnetic signals. *IEEE Transactions on Computers, 69*(3), 312-326.

Sethi, K., Kumar, R., Sethi, L., Bera, P., & Patra, P. K. (2019). *A novel machine learning based malware detection and classification framework.* Paper presented at the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security).

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. *IEEE access, 8*, 222310-222354.

Sreekumari, P. (2020). *Malware detection techniques based on deep learning.* Paper presented at the 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS).

Ullah, F., Naeem, H., Jabbar, S., Khalid, S., Latif, M. A., Al-Turjman, F., & Mostarda, L. (2019). Cyber security threats detection in internet of things using deep learning approach. *IEEE access, 7*, 124379-124389.

Vinayakumar, R., Alazab, M., Soman, K., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE access, 7*, 41525-41550.

Vinayakumar, R., Alazab, M., Soman, K., Poornachandran, P., & Venkatraman, S. (2019). Robust intelligent malware detection using deep learning. *IEEE access, 7*, 46717-46738.

Xu, S., Xia, Y., & Shen, H.-L. (2020). Analysis of malware-induced cyber attacks in cyber-physical power systems. *IEEE Transactions on Circuits and Systems II: Express Briefs, 67*(12), 3482-3486.

Zhang, F., Kodituwakku, H. A. D. E., Hines, J. W., & Coble, J. (2019). Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data. *IEEE Transactions on Industrial Informatics, 15*(7), 4362-4369.

Zhu, J., Jang-Jaccard, J., & Watters, P. A. (2020). Multi-loss Siamese neural network with batch normalization layer for malware detection. *IEEE access*, *8*, 171542-171550.

Foundation University Journal of Engineering and Applied Sciences
Vol. 3, Issue 1, 2022.
DOI: 10.33897/fujeas.v3i1.506

Research Article

Facial Based Gender Classification for Real Time Applications

Anmol Haider[1], Muhammad Imran[1*]

[1]Shaheed Zulfiqar Ali Bhutto Institute of Science and Information Technology, Islamabad, Pakistan

[*]Corresponding Author: Muhammad Imran. Email: dr.imran@szabist-isb.edu.pk

Abstract:

Appearance and facial features play an important role in gender recognition through images. For gender classification, multiple techniques were presented to acquire better results in which preprocessing part is one of the major and very important for gender classification as it removes noise, enhances, images, and eliminates any unnatural colors from an image. Another major aspect is the efficient feature extraction method. If features extracted accurately then the result of classification will improve. Over the past few years, gender classification techniques work perfectly for a controlled environment. However, challenges occurred for real-time applications due to low resolution, off-angle poses, faces with occlusion, and various expressions. The main focus of this study is to overcome existing challenges and propose a method that can be implemented in real-time applications. This research work proposed a novel method in which CNN has been used for classification of gender for real-time application. To assess the performance of proposed method experiments were conducted on static images and video data sets. The proposed research work achieved 98% of accuracy during the experiments.

Keywords: Gender classification; Recognition; Feature extraction; real-time application

1. Introduction

The human face conveys a lot of information that is easy to identify by humans, however, it is difficult to identify by machines. In recent years, gender classification attracts the attention of various researchers (Lin, Wu, Zhuang, Long, & Xu, 2016; Ng, Tay, & Goi, 2015). Different gender classification techniques have been proposed by researchers in the last few decades (Abbas et al., 2021; Cartwright & Nancarrow, 2022; Lin et al., 2016; Moeini & Mozaffari, 2017; Ng et al., 2015).

Now a day's deep learning is a latest technology just likes driverless cars. It is also key of voice control in consumer's devices such as tablets, TVs, Phones speakers and hand frees. There are a number of deep learning networks used for different purpose such as convolutional neural network, Generative adversarial network, Wasserstein convolutional neural network, style transfer network and so on. For deep learning we first learn the models that perform specific task including understanding of forms texts, images, digital images, black images and sound. Deep learning model can give good high accuracy while some times its performance is exceeding the human levels. In deep learning the model is trained by using labeled data, images, and architecture of the neural network contain numbers of CNN layers. Deep learning is used for several computer vision applications such as image classification, object detection, object segmentation, image reconstruction, image super resolution and image style transfer. It is considered most important in the fields of the Surveillance system, Security purposes, Mobile applications, Advertisements, and many more. With so many applications there are still various challenges present such as analysis of automatic video data is a very difficult task, face detection through a live stream is also a difficult task due to various expressions, different poses, face alignment, illumination conditions. Various researchers presented their solution to overcome these problems.

In recent years, gender classification attracts various researchers due to its usage in almost every field such as attendance system, surveillance system, security purposes, mobile applications, advertisements, and many more.

However, users' need a system that can be applied in real-time applications because of increasing development in the field of the internet nowadays, live video has much attraction with many users that can share their videos. These shared videos from real-world record human faces that's why analysis of face is very important in real-time applications.

Many methods have already been proposed for gender classification in both controlled and uncontrolled situations. However, problems occurred in an uncontrolled situation when there are a high rate of noises, various illumination conditions, and occluded faces or covered faces in real-time applications (Lin et al., 2016; Ng et al., 2015). To mitigate these problems, this research work proposed a new method that will work on uncontrolled conditions and enhances the performance of gender classification techniques in real-time application.

The remaining sections of the paper are as follows. Section 2 presents the literature review, and Section 3 Section 3 discusses the proposed framework. Results and discussed in Section 4. Conclusion is presented in Section 5.

2. Literature Review

Feature extraction is one of the most important steps in gender classification. Bukar et al. (Bukar, Ugail, & Connah, 2016) proposed a method SAM (supervised appearance model). The proposed method addressed the problem of feature extraction for gender classification. In the recent past, the Active Appearance Model (AAM) was used to capture the shape and texture variation for feature extraction.

AAM utilized the Principle Component Analysis (PCA) for a dimensional reduction in an unsupervised manner but cannot handle how the predictor variables related class labels mean it's only used to detect the texture of face image. To overcome this problem authors proposed a model named as Supervised Appearance Model (SAM) which replace PCA with PLS (Partial Least Square).

The proposed method is performed by forming a parameterized model using PLS dimensionality reduction to capture the variations as well as combine them in a single model. PLS can do both dimensionality reduction and regression simultaneously. The results of the experiment were compared with previous well know techniques.

Antipov et al. (Antipov, Berrani, & Dugelay, 2016) presented an approach named DCNN (Deep Conventional neural network) which is an advanced form of convolution neural network (CNN) that is a very powerful recognition technique having different layers in its architecture. Face detection was performed by voila Jones

Experiments were performed on LFW and CASIA web face databases Images in both databases were centered faces having resolution 250*250.

Zhang et al. (Zhang & Xu, 2018) proposed the Local deep neural network (LDNN) technique. In this proposed model, local image patches were selected based on the detected facial landmarks. The detected patches then used for network training which reduced the cost of training Author proved that local deep neural network (LDNN) was cheaper in terms of computational time than (Conventional neural network) CNN. The Face detection was obtained by viola Jones. The experiments were performed on the Audience dataset which contained 26580 images of the face. The proposed framework achieved 80.64% accuracy.

Briones et al. (González-Briones, Villarrubia, De Paz, & Corchado, 2018) presented a method named a multi-agent system. In the proposed method multiple classifiers were applied for the experiment and compared each other to obtain the best classifier. The proposed method was composed of well-known techniques such as Fisher Faces, and LBP (Local Binary Pattern) for face recognition. Face recognition was performed through a combination of fisher faces and ANN (Artificial neural network). LPB (local binary pattern) was used as a feature extraction technique. A bilateral filter was used to extract edges of faces. Experiments were performed on the FERET database having 14051 images with different angles. The result of experiments achieved 85% accuracy.

Santana et al. (Castrillón-Santana, Lorenzo-Navarro, & Ramón-Balmaseda, 2016) presented a multi-scale approach where features were extracted from the face, head, and shoulders areas. This method used various feature extraction techniques such as HOG (Histogram of Oriented Gradients), LBP (Local Binary Patterns), LTP (Local Ternary Patterns), and WLD (Weber Local Descriptor) LOSIB (Local Oriented Statistics Information Booster).

Mansanet et al. (Mansanet, Albiol, & Paredes, 2016) presented a method named Local Deep Neural Network (LDNN). In the proposed method Local-DNN was responsible to obtain local features. This proposed method was the general framework that applies only in the local feature. SVM (Support Vector Machine) was adopted as a classifier. Experiments were performed on two databases such as (LFW) Labeled Faces in the Wild contained 13233 face images and the Gallagher database contained 28231 face images. Proposed framework achieved 96.25% on LFW database whereas 90.50% on Gallagher's Database.

Huang et al. (Huang et al., 2014) proposed a method named Local circular pattern (LCP). In the past Local binary patterns (LBP) were used to extract the features. However, LBP still has various limitations such as it cannot deal with noisy images because it worked on image pixel values. To overcome this problem LCP was proposed, that worked on Cluster-based quantization rather than binary quantization hence this easily remove noise. Experiments were performed on the FRGC database having 1876 images. And achieved 95.65 accuracy. Off angle face images are still a very challenging field for recognition of a face.

Alomar et al. (Alomar et al., 2013) proposed a multi-scale Band let and local binary pattern (LBP) method for gender classification from face images. The band let is one of the multi-resolution techniques that can adapt to the orientation of edges, and also it enabled better capturing of texture from face images. In the proposed method LBP and Band let were used to extract the features and minimum distance classifier (MDC) was utilized to classify the gender. The first Band let transformation was performed to detect the geometric shape of the image. After that LBP was applied to extract the features of the face image. The experiments were performed using FERET grayscale face database having 994 images and achieved 95.8%accuracy.

Alignment of the face is a very common problem and to solve this problem Eidinger et al. (Eidinger, Enbar, & Hassner, 2014) proposed a method in which three steps were performed. The first step was Detection of the face, which was performed by voila & Jones. The second step was feature extraction and LBP was used for it. After that gender classification was performed by dropout SVM that able to remove the over fitting problem. The experiment was performed on the Gallagher database which contained 28231 images and achieved 88.6% accuracy.

Rai et al. (Rai & Khanna, 2014) worked on the front face images by the use of feature selection which is based on combine information and fusion of extracted features from shape and texture of face images for classification of gender.

Experiments were performed on the FERET database containing more than 14 thousand images and only

380 images used for testing. The result of the experiment achieved 78.71% accuracy.

Perez et al. (Perez, Tapia, Estévez, & Held, 2012) presented a method for gender classification named a local binary pattern (LBP) based classifier. In the proposed method LBP was used to for feature extraction named MCT (Modified Census Transform) which was responsible for extracting the feature of the face twice to enhance the performance of the face extraction process. The proposed method was used to extract the local feature therefore face alignment was also an important achievement for this approach. Face alignment used the location of fiducial points such as eyes, mouth, and nose so on. Experiments were performed on LWF (Labeled wild faces) database containing 13,233 images. Adaboost was used as a classifier which gives 87%accuracy which was higher as compared to another classifier such as Gabor, jets.

3. Proposed Framework

3.1 Overview

The flowchart of the proposed method for gender classification is shown in Figure 1. The first step is the face detection step, Cascade face detection model is used for face detection in live video. The next step is preprocessing, in this step, images are feed into BLOB which is used to enhance the quality of an image. Finally, classification is performed by CNN.

The details of these steps are mentioned below. In live video streaming gender, classification is a challenging field, and it is important for real-time applications but due to occluded faces, blur motion, various illumination conditions these mentioned problems still exist. This research is focused on the above-mentioned challenges that occurred during the face detection phase for gender classification in real time application.

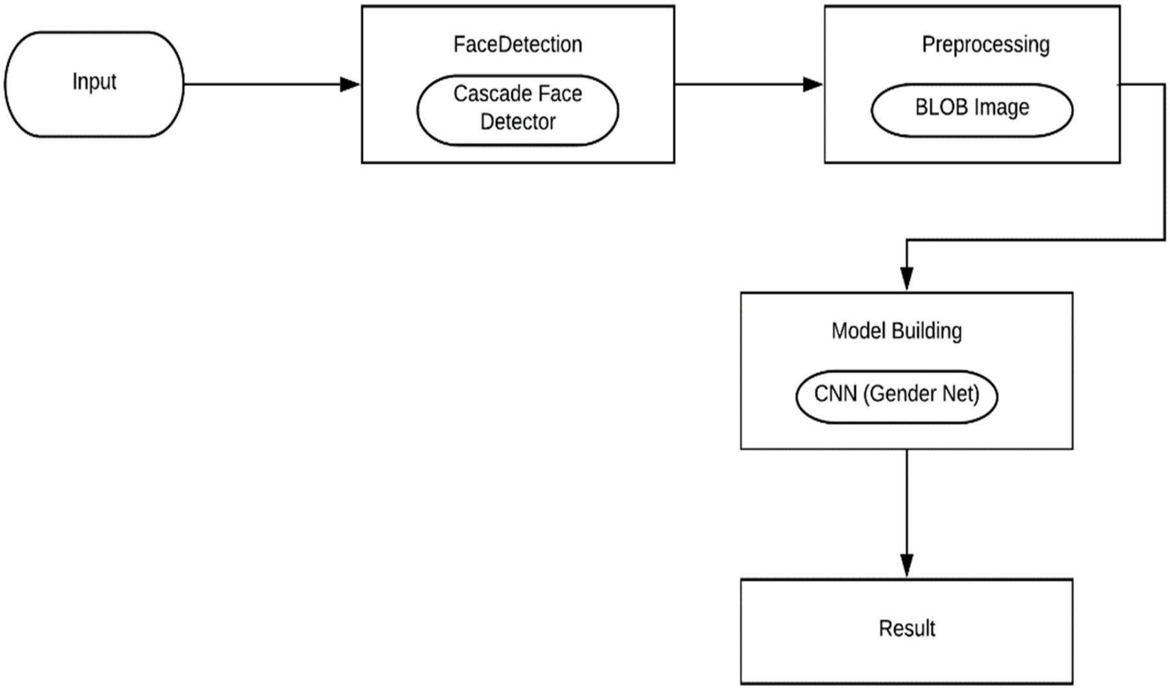


Figure 1: Proposed framework

3.2 Input

Classification of gender is performed by three type of input such as multiple faces are detected from video. Static images are also detected by this proposed method. Real time gender classification is also performed by given method.

3.3 Face Detection

As input data can have multiple objects so first step is face detection from input data. In the proposed technique Face is detected using Haar Cascade. The Haar cascade extract various feature such as lines, edges, and rectangles. An edge is explained as a sharp change in contrast from bright to dark in either vertical or horizontal direction so whenever the Haar cascade algorithm is applied on the square of pixels it follows that property, it marks them as an edge feature. The whole region of the image is extracted for lines, edges, and rectangles. We keep the rectangle on the face for not losing of the face because in cascade classifier one single blink of an eye is also responsible to lose a face, so we adjust the rectangle according to the face position.

3.4 Preprocessing

After detecting the face next step is preprocessing. In this step detected face is converted into the BLOB (Binary Large Objects). BLOB is used to enhance the quality of image such as illumination condition and normalization by using of mean subtraction and scale factor. It has four steps such as resizes and crop images from the middle of subtracting mean values, scales values by scale vector, and swap Blue (B) and Red (R) channels.

3.4.1 Mean Subtraction

Illumination is the amount of source light that destroy the pixel value Mean subtraction is applied when there were illumination changes in the input images. Therefore, it is used to aid in Convolution Neural Networks as a technique.

Typically, the results of these three topple consist of the mean value of the Red, Green, and Blue channels respectively. In other cases, the mean values of Red, Blue, and Green is computed channel wise other than pixel-wise. Both these methods are perfectly valid forms of mean subtraction. When the image is ready to pass our proposed system, we subtract the mean from every input channel of the input image.

$$R = R - U_R \tag{1}$$

$$G = G - U_G \tag{2}$$

$$B = B - U_B \tag{3}$$

3.4.2 Scale Factor

Image scaling is used to resize the images. The scaling factor is used in the proposed method which adds in the normalization of the image. As mentioned above the scaling is used to normalize the image so that all images should be in same size. Scaling is done for Red, Green and Blue component of the image.

$$R = (R - U_R)/\sigma \tag{4}$$

$$G = (G - U_G)/\sigma \tag{5}$$

$$B = (B - U_B)/\sigma \tag{6}$$

We also manually set the scale factor to scale the input image space into a particular range.

3.5. Model Building

The CNN performs both feature extraction and classification within a single network structure through learning on data samples. CNN is specifically designed to cope with shortcomings of the traditional feature extractor that is characterized by being static, is designed independently of the trainable classifier, and is not part of the training procedure. A final benefit of CNNs is that they are relatively easier to train since they have fewer parameters than fully connected MLP neural networks with the same number of hidden layers. Filters are used to analyze the value of nearby pixels. By the rule of thumb take 5 * 5 filter size then move it on the image from upper left to lower right. For every point on the image, the filter value is calculated by using the convolution operation. Filters reduce the number of weights in the neural network when the location of features such as eyes, nose, lips, and cheeks changes then classification is not performed by a neural network. When a model building starts, we manually set the values of filters after that is continuously updated throughout the training process.

Convolution is composed of independent filters. Every filter is independently convolved with image and ends with six feature maps. The pooling layer is used to reduce the spatial size of the image and it operates on each feature map separately. Block diagram of CNN is shown in Figure 2.

4. Experimental Results

4.1 Performance Parameters

To assess the performance of proposed solution following established performance parameters are used in which precision determine the number of positive class means how many times proposed method gives accurate result.

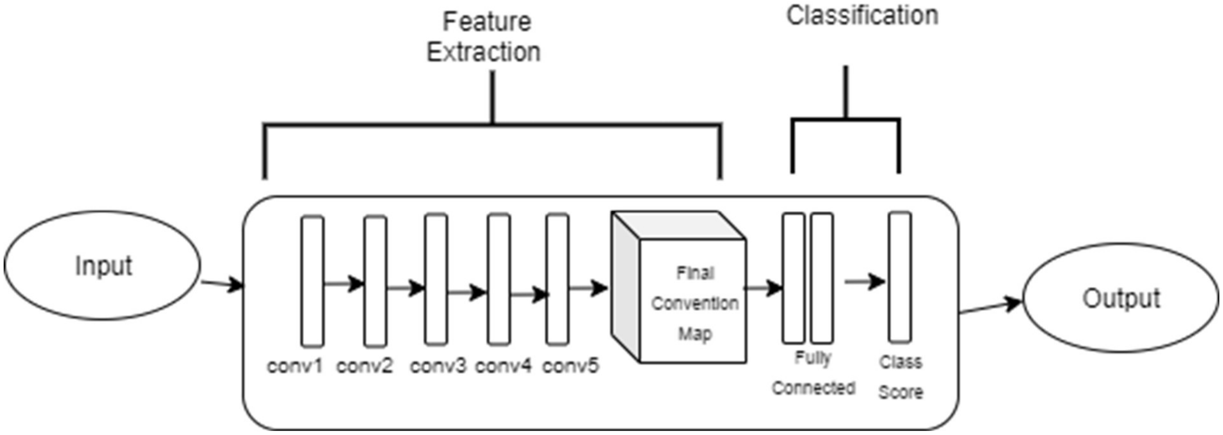


Figure 2: Block diagram of CNN

4.1.1 Precision

Precision is the percentage of the relevant result that is taken from the system during training or testing. In this research, Precision determines the number of positive class predictions that belong to the positive class only.

$$Precision = True\ Positive \div (ture\ positive + flase\ positive) \quad (7)$$

4.1.2 Recall

The recall is the percentage of relevant result that is classified by the proposed method in this research. The recall is determining the number of positive class predictions made from all positive.

$$Recall = True\ Positive \div (positive + Flase\ Negative) \quad (8)$$

4.2 Comparative Analysis of Experimental Study I (Static Images) Avg

To validate the performance of the proposed system, the results of the proposed systems are compared with various other well-known techniques. These techniques include Santana et al (Castrillón-Santana et al., 2016) , Endanger et al. (Eidinger et al., 2014), and Xu et al. (Zhang & Xu, 2018). The results are shown in Table 1 and Figure 3.

Table 1: Comparison table of static images

Methods	**Accuracy**	**Precision**	**Recall**	**F-measure**
Santana et al. (Castrillón-Santana et al., 2016), 2016	92.46	0.897	0.896	0.896
Endanger et al. (Eidinger et al., 2014), 2014	88.6	0.815	0.810	0.812
Zhang et al. (Zhang & Xu, 2018), 2018	93	0.901	0.902	0.901
Proposed Method	98	0.95	0.94	0.96

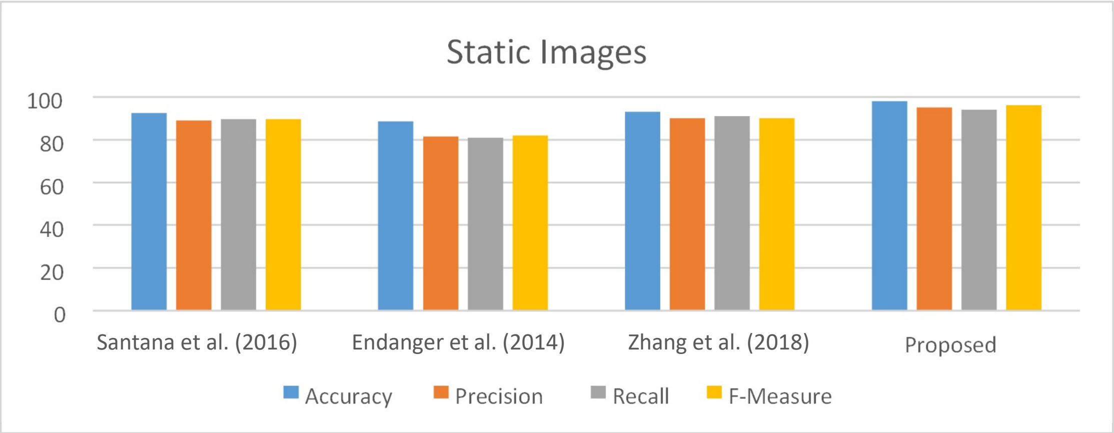


Figure 3: Comparison chart of static images

4.3 Comparative Analysis of Experimental Study 2 (Videos)

To validate the performance of the proposed system, the results of the proposed systems are compared with various other well-known techniques. These techniques include Bukar et al. (Bukar et al., 2016), Mansanet et al. (Mansanet et al., 2016), and Perez et al. (Perez et al., 2012). A comparative analysis is shown in Table 2 and Figure 4.

Table 2 Comparison table of Videos

Methods	Accuracy	Precision	Recall	F-measur
Bukar et al. (Bukar et al., 2016) 2016	92.50	0.897	0.896	0.896
Mansanet et al. (Mansanet, Albiol, & Paredes, 2016)	96.25	0.95	0.91	0.92
Perez et al. (Perez et al., 2012)	87	0.85	0.86	0.82
Proposed Method	98.1	0.905	0.901	0.905

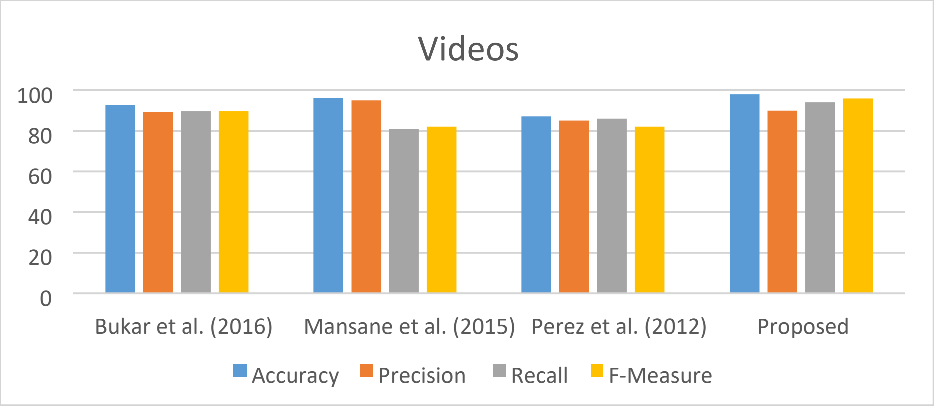


Figure 4: Comparison chart of videos

From Table 2 and Figure 4, one can observe that performance of the proposed approach is better than the previous approaches. As for accuracy, precision, and recall and F-measure of the proposed method is higher than the previous research.

4.3 Comparative Analysis of Experimental Study 3 (Real-Time)

To validate the performance of the proposed system in real-time, the results of the proposed systems are compared with various other well-known techniques. These techniques include Bukar et al. (Bukar et al., 2016), Mansanet et al. (Mansanet et al., 2016), Comparative analysis is shown in Table 3 and Figure 5.

From Table 3 and Figure 5, it is observed that performance parameter such as accuracy, precision-recall and F-measure of the proposed method is higher than the previous research.

Table 3: Comparison table of Real-Time

Methods	Accuracy	Precision	Recall	F-measure
Bukar et al., (Bukar et al., 2016), 2016	92.50	0.897	0.896	0.896
Mansanet et al., (Mansanet et al., 2016), 2016	96.25	0.95	0.91	0.92
Zhang et al., (Zhang & Xu, 2018), 2018	87	0.85	0.86	0.82
Proposed Method	98.1	0.905	0.901	0.905

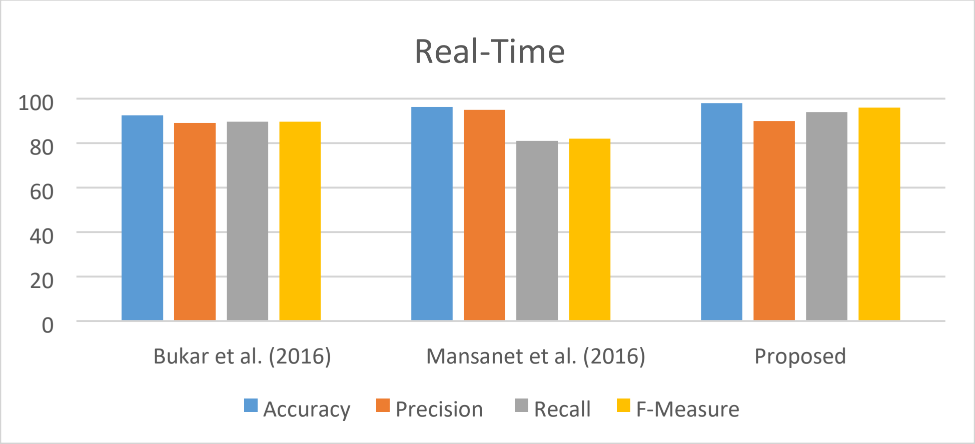


Figure 5: Comparison chart of real-time

5. Conclusion

In this study, the framework is proposed for the classification of gender in real time application. Classification of gender is very important in surveillance systems, security purposes, mobile applications, advertisements, and many more.

It is a very challenging field of image processing and many researchers have presented their work to overcome limitations such as low-quality images, illumination conditions, and live video face detection in real-time. This research compares different classification methods that have been proposed by various researchers. After considering current literature there are still various problems such as only a few researchers considered enhancement techniques in their approach and real-time applicability is still a major problem in current methods. This research proposed a method that can overcome current literature challenges such as face alignment, illumination conditions, and real-time applicability. The proposed method is used for three types of input from images, videos, and real-time applications.

LWF, Groups, and GCLV three databases are used to apply proposed techniques to verify the accuracy and we achieved 99% accuracy in real-time gender detection which is higher accuracy as compared to previous research. In the preprocessing, step BLOB images are used to enhance the illumination condition and image quality. We apply the CNN model after the BLOB images to classify gender and feature extraction as well. In this research three types of input are taken which is a static image, multiple face detection in video, and Real-time gender detection. No researcher has done work on all three inputs.

References

Abbas, F., Yasmin, M., Fayyaz, M., Abd Elaziz, M., Lu, S., & El-Latif, A. A. A. (2021). Gender classification using proposed CNN-based model and ant colony optimization. *Mathematics, 9*(19), 2499.

Alomar, F. A., Muhammad, G., Aboalsamh, H., Hussain, M., Mirza, A. M., & Bebis, G. (2013). *Gender recognition from faces using bandlet and local binary patterns.* Paper presented at the 2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP).

Antipov, G., Berrani, S.-A., & Dugelay, J.-L. (2016). Minimalistic CNN-based ensemble model for gender prediction from face images. *Pattern recognition letters, 70*, 59-65.

Bukar, A. M., Ugail, H., & Connah, D. (2016). Automatic age and gender classification using supervised appearance model. *Journal of Electronic Imaging, 25*(6), 061605-061605.

Cartwright, T., & Nancarrow, C. (2022). A Question of Gender: Gender classification in international research. *International Journal of Market Research, 64*(5), 575-593.

Castrillón-Santana, M., Lorenzo-Navarro, J., & Ramón-Balmaseda, E. (2016). On using periocular biometric for gender classification in the wild. *Pattern recognition letters, 82*, 181-189.

Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security, 9*(12), 2170-2179.

González-Briones, A., Villarrubia, G., De Paz, J. F., & Corchado, J. M. (2018). A multi-agent system for the classification of gender and age from images. *Computer Vision and Image Understanding, 172*, 98-106.

Huang, D., Ding, H., Wang, C., Wang, Y., Zhang, G., & Chen, L. (2014). Local circular patterns for multi-modal facial gender and ethnicity classification. *Image and Vision Computing, 32*(12), 1181-1193.

Lin, F., Wu, Y., Zhuang, Y., Long, X., & Xu, W. (2016). Human gender classification: a review. *International Journal of Biometrics, 8*(3-4), 275-300.

Mansanet, J., Albiol, A., & Paredes, R. (2016). Local deep neural networks for gender recognition. *Pattern recognition letters, 70*, 80-86.

Moeini, H., & Mozaffari, S. (2017). Gender dictionary learning for gender classification. *Journal of Visual Communication and Image Representation, 42*, 1-13.

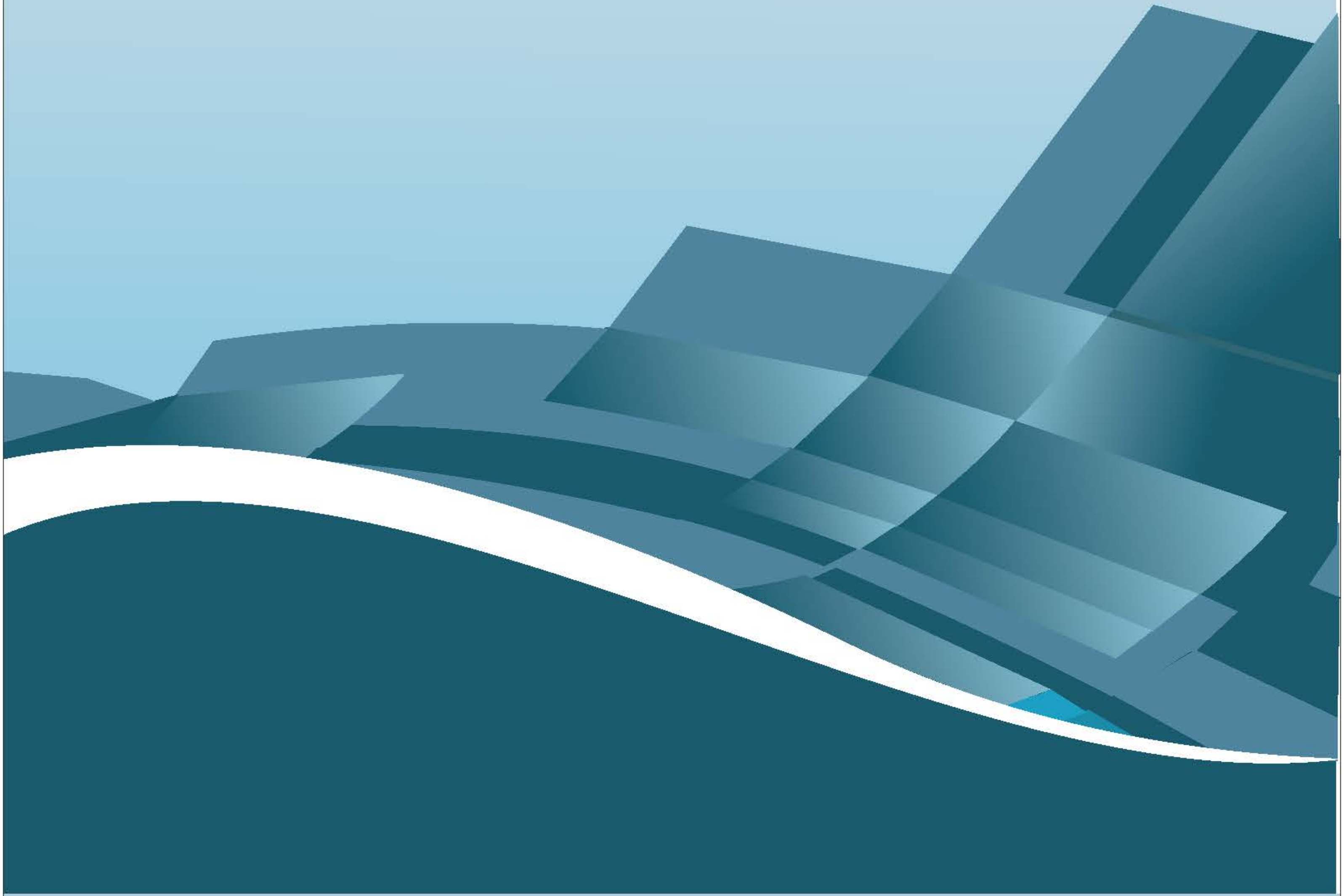
Ng, C.-B., Tay, Y.-H., & Goi, B.-M. (2015). A review of facial gender recognition. *Pattern Analysis and Applications, 18*, 739-755.

Perez, C., Tapia, J., Estévez, P., & Held, C. (2012). Gender classification from face images using mutual information and feature fusion. *International Journal of Optomechatronics, 6*(1), 92-119.

Rai, P., & Khanna, P. (2014). A gender classification system robust to occlusion using Gabor features based (2D) 2PCA. *Journal of Visual Communication and Image Representation, 25*(5), 1118-1129.

Zhang, Y., & Xu, T. (2018). Landmark-guided local deep neural networks for age and gender classification. *Journal of Sensors, 2018*.

Foundation University Journal of Engineering and Applied Sciences



The Secretariat FUJEAS, Foundation University School of Science & Technology (FUSST)
New Lalazar, Rawalpindi.
info.fujeas@fui.edu.pk, https://fujeas.fui.edu.pk/, +92-51-5151437-38 (Ext. 300)