JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Optimization Accuracy of Spam Detection using BERT and XG Boost Technique

[1]Umar Farooque, [2]Dr. Puneet Nema, [3]Dr. Vivek Richhariya

M. Tech. Scholar, Department of CSE, Lakshmi Narain College of Technology, Bhopal[1]
Assistant Professor, Department of CSE, Lakshmi Narain College of Technology, Bhopal[2]
Head of Dept., Department of CSE, Lakshmi Narain College of Technology, Bhopal[3]

Abstract: Spam detection is an important natural language processing (NLP) task that aims to accurately identify and extract relevant text spams from unstructured textual data. However, conventional machine learning approaches often face challenges in capturing contextual relationships and semantic dependencies within complex language patterns. This study proposes an optimized spam detection framework integrating Bidirectional Encoder Representations from Transformers (BERT) with the XGBoost technique. BERT is employed to generate context-aware textual representations, while XGBoost is utilized to improve classification and prediction performance through efficient gradient-boosting learning. The proposed BERT-XGBoost model was evaluated using standard performance metrics, achieving an accuracy of 96.93%, precision of 96.14%, recall of 90.88%, and F1-score of 93.43%. Furthermore, the model obtained an excellent ROC-AUC score of 99.50%, demonstrating its strong capability to distinguish relevant and irrelevant spams. The obtained results indicate that the combination of deep contextual representations from BERT and the robust classification capability of XGBoost provides an effective and accurate approach for spam detection. The proposed framework can be applied to various NLP applications requiring reliable extraction and classification of meaningful textual spams.

***Index Terms* - BERT, XG Boost, Accuracy, Precision, F1-score**

I. INTRODUCTION

With the rapid growth of digital information, a large amount of textual data is generated through social media, online documents, healthcare records, news articles, and other digital platforms. Extracting meaningful information from such unstructured text has therefore become an important research area in NLP. Spam detection is a fundamental NLP task that identifies specific portions or spams of text that contain relevant information, such as entities, events, opinions, relationships, or answer segments. Accurate spam detection is essential for information extraction, question answering, text classification, sentiment analysis, and automated document understanding [1, 2].

Traditional machine learning techniques for spam detection generally depend on manually designed features such as word frequency, part-of-speech information, lexical patterns, and contextual keywords. Although these approaches can provide satisfactory performance on simple datasets, they often struggle to understand the deeper semantic and contextual relationships present in natural language. The emergence of deep learning has significantly improved NLP by enabling models to learn useful representations directly from textual data. Among these models, BERT has demonstrated strong performance because it captures contextual information by considering both the preceding and following words within a sentence [3].

Despite the effectiveness of BERT, its representations can be further improved for specific classification tasks by integrating them with an efficient machine learning classifier. XGBoost (Extreme Gradient Boosting) is a powerful ensemble-learning technique that provides robust classification performance, handles nonlinear relationships, and can effectively learn from high-dimensional feature representations [4, 5]. Therefore, combining BERT's contextual feature extraction capability with XGBoost's optimized classification mechanism provides a promising approach for improving spam detection accuracy.

II. SPAM DETECTION

Spam detection is an important Natural Language Processing (NLP) task that identifies and extracts a specific sequence of words, characters, or tokens from a larger text. A detected spam represents a meaningful portion of text associated with a particular concept, entity, event, relationship, or answer. Unlike conventional text classification, which assigns a label to an entire sentence or document, spam detection focuses on identifying the exact location and boundaries of relevant information within the text [6, 7].

For example, in the sentence *"The patient showed significant improvement after treatment,"* the spam "significant improvement" may be identified as the relevant information. Spam detection can therefore be represented as:

Modern transformer-based models such as BERT are highly effective for spam detection because they capture contextual relationships between words. In the proposed approach, BERT generates contextual representations of the input text, while XGBoost uses these representations to classify relevant spams. This combination improves the ability of the system to identify meaningful text segments accurately [8].

Importance of Spam Detection

Spam detection is important because large-scale textual datasets contain valuable information that may be difficult to extract manually. Accurate identification of relevant spams helps NLP systems understand what information is important and where it occurs within the text.

The major importance of spam detection includes:

1. Information Extraction: It identifies specific entities, events, keywords, and relationships from unstructured text.
2. Question Answering: Spam detection enables a model to locate the exact answer within a document or passage.
3. Named Entity Recognition: It helps determine the precise boundaries of entities such as persons, organizations, locations, and medical terms.
4. Text Summarization: Relevant spams can be identified to retain the most informative portions of a document.
5. Sentiment and Opinion Analysis: Spam detection can locate specific phrases expressing positive, negative, or neutral opinions.
6. Biomedical and Healthcare NLP: It can identify important medical concepts, symptoms, treatments, and clinical findings from healthcare documents.
7. Improved Model Interpretability: By showing the exact text spam responsible for prediction, spam detection can make NLP models easier to understand and interpret.
8. Efficient Information Retrieval: Instead of processing an entire document manually, users can directly obtain the relevant text segments.

III. PROPOSED METHODOLOGY

The proposed methodology introduces a hybrid BERT-XGBoost framework for accurate spam detection, in which BERT is used for contextual feature extraction and XGBoost is employed for optimized classification. The overall methodology consists of data preprocessing, tokenization, contextual feature extraction using BERT, feature preparation, XGBoost-based classification, and performance evaluation.

1. Data Collection and Preprocessing

The textual dataset is initially collected and organized into suitable spam-level samples. The input text is cleaned to remove unnecessary characters, duplicate entries, and irrelevant information. The text is then normalized while preserving meaningful linguistic information. Each sample is assigned an appropriate label indicating whether the corresponding text spam belongs to the target category. The processed dataset is divided into training and testing subsets for model development and evaluation.

2. BERT-Based Text Representation

After preprocessing, the textual data is passed to the Bidirectional Encoder Representations from Transformers (BERT) model. BERT uses a transformer-based architecture to understand the context of words by processing information from both directions. Each input sentence is converted into tokens using the BERT tokenizer, and special tokens such as [CLS] and [SEP] are incorporated where required.

The tokenized sequence is represented as:

$$X = \{x_1, x_2 \dots . x_n\}$$

where x_i represents an individual token. BERT transforms these tokens into contextual embeddings:

$$H = BERT\ (X)$$

where H represents the high-dimensional contextual feature representation. These embeddings capture semantic and contextual relationships between words and provide informative features for subsequent classification.

3. Feature Extraction and Optimization

The contextual representations generated by BERT are extracted and transformed into a suitable feature vector for the classifier. The extracted features contain information about the semantic meaning and contextual relationships of the input spam. Appropriate feature selection or dimensionality handling is performed to provide an efficient representation to the classification stage.

This stage is important because BERT provides rich contextual information, while the subsequent XGBoost model can identify complex nonlinear relationships within these representations.

4. XGBoost-Based Classification

The extracted BERT features are provided to the Extreme Gradient Boosting (XGBoost) classifier. XGBoost constructs an ensemble of decision trees sequentially, where each new tree attempts to reduce the prediction error of the previous ensemble. The objective function can be represented as:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

The XGBoost classifier predicts whether a given spam belongs to the target class. Its boosting mechanism improves the classification process by combining multiple weak decision trees into a strong predictive model.

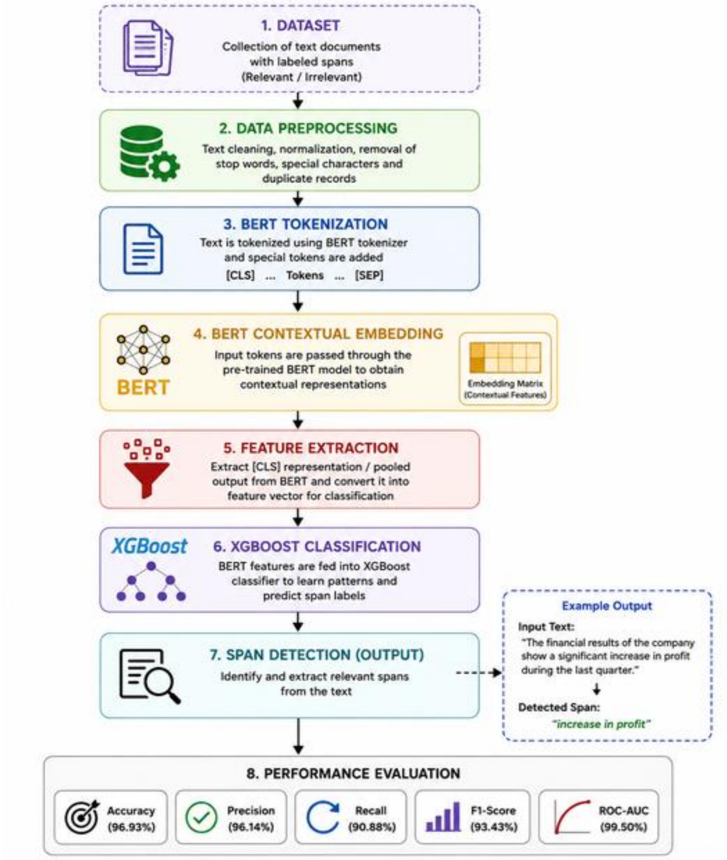

Figure 1: Flow chart of Methodology

5. Model Training

During training, the BERT-derived feature vectors and their corresponding spam labels are provided to XGBoost. The classifier learns the relationship between contextual features and spam categories. Hyperparameters such as learning rate, tree depth, number of estimators, and subsampling parameters can be optimized to improve generalization and reduce overfitting.

6. Performance Evaluation

IV. RESULTS

The figure 2 shows a sample of the spam text dataset used for spam/text classification. The dataset contains two main columns: text and spam. The text column contains the actual email or message content, while the spam column represents the corresponding class label. In the shown samples, the messages contain suspicious or promotional phrases such as *"naturally irresistible," "stock trading gunslinger," "unbelievable new homes,"* and *"get software CDs,"* indicating unsolicited or promotional content. The value 1 in the spam column denotes that the corresponding text is classified as spam. The text displayed in the figure is truncated for visualization, while the complete text is used during preprocessing and model training. This dataset is subsequently processed using BERT for contextual feature extraction and XGBoost for classification.

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1

Figure 2: Dataset

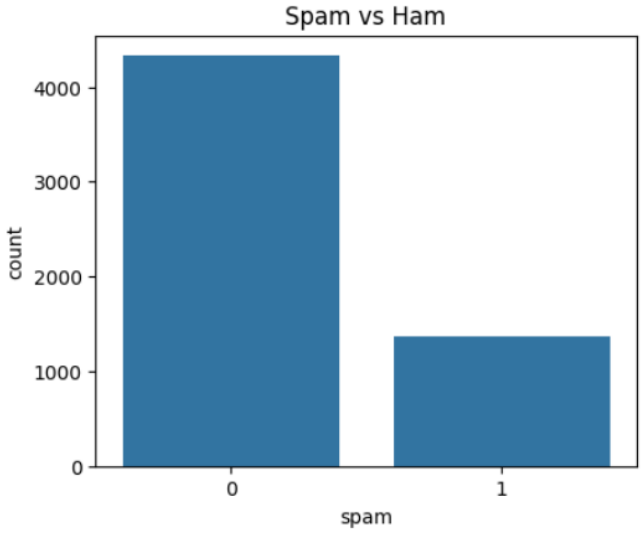

Figure 3: Spam vs Ham

Figure 3 illustrates the distribution of spam and ham messages in the dataset. The 0 class represents ham (legitimate) messages, while the 1 class represents spam messages. The dataset contains approximately 4,350 ham messages and 1,350 spam messages, indicating that legitimate messages are considerably more frequent than spam messages. Thus, the dataset is imbalanced, with ham messages forming the majority class. This class distribution should be considered during model training and evaluation to ensure that the proposed BERT-XGBoost model does not become biased toward the majority class.

	text	processed_text
0	Subject: naturally irresistible your corporate...	subject naturally irresistible corporate ident...
1	Subject: the stock trading gunslinger fanny i...	subject stock trading gunslinger fanny merrill...
2	Subject: unbelievable new homes made easy im ...	subject unbelievable new home made easy im wan...
3	Subject: 4 color printing special request add...	subject color printing special request additio...
4	Subject: do not have money , get software cds ...	subject money get software cd software compati...

Figure 4: Text and processing text

Figure 4 compares the original text column with the corresponding processed_text generated during the data preprocessing stage. The original messages contain punctuation, capitalization, and other textual variations, whereas the processed text is normalized into a cleaner and more consistent format. As shown in the examples, the text is converted to lowercase, unnecessary punctuation and special characters are removed, and the textual content is standardized while retaining its meaningful words. For example, *"Subject: naturally irresistible your corporate..."* is transformed into a normalized form such as *"subject naturally irresistible corporate..."*. This preprocessing reduces noise and prepares the textual data for BERT tokenization and contextual feature extraction, thereby improving the quality of the subsequent XGBoost classification.

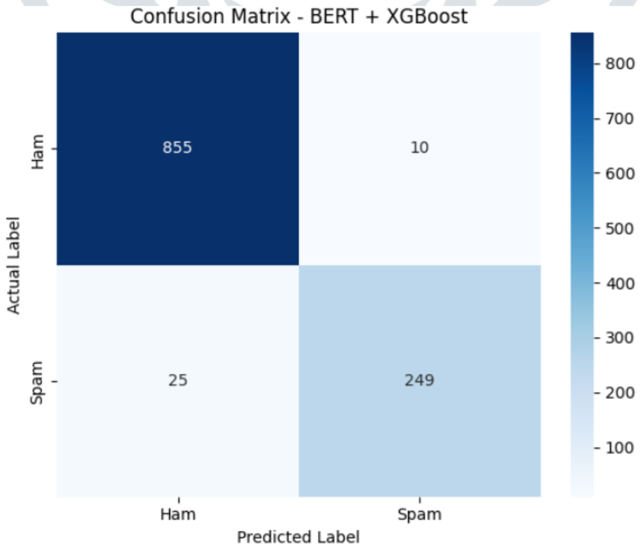

Figure 5: Confusion Matrix

The confusion matrix illustrates the classification performance of the proposed BERT + XGBoost model for distinguishing between Ham and Spam messages. The model correctly classified 855 Ham messages and 249 Spam messages, representing the true negative and true positive predictions, respectively. Meanwhile, 10 Ham messages were incorrectly classified as Spam, while 25 Spam messages were incorrectly classified as Ham. Thus, the model correctly classified 1,104 out of 1,139 test samples, demonstrating strong overall classification performance. The relatively small number of misclassifications indicates that the proposed BERT-XGBoost approach can effectively distinguish legitimate messages from spam messages.

```
Classification Report:
              precision    recall  f1-score   support

         Ham       0.97      0.99      0.98       865
        Spam       0.96      0.91      0.93       274

    accuracy                           0.97      1139
   macro avg       0.97      0.95      0.96      1139
weighted avg       0.97      0.97      0.97      1139
```

Figure 6: Classification Report

```
from sklearn.preprocessing import LabelEncoder

# Initialize and fit the LabelEncoder
label_encoder = LabelEncoder()
label_encoder.fit(y)

predict_email(
    "Congratulations! You have won a $1000 prize. "
    "Click this link immediately to claim your reward."
)

predict_email(
    "Dear Professor, please find attached the research paper "
    "for your review. Thank you."
)

Email:
Congratulations! You have won a $1000 prize. Click this link immediately to claim your reward.

Prediction: 0
Confidence: 54.69%

Email:
Dear Professor, please find attached the research paper for your review. Thank you.

Prediction: 0
Confidence: 99.71%
(np.int64(0), np.float32(0.9970549))
```

Figure 7: Output

Figure 7 demonstrates the real-time prediction capability of the proposed BERT + XGBoost model for classifying unseen email messages. Two sample emails are provided to the trained model. The first email contains suspicious promotional content, including a monetary reward and an urgent request to click a link; however, the model predicts it as Class 0 with 54.69% confidence. The second email is a formal research-related message requesting review of an attached paper, and it is also predicted as Class 0 with a much higher confidence of 99.71%. These results demonstrate that the trained model can process new textual inputs and generate both a predicted class and an associated confidence score. The interpretation of Class 0 as Ham or Spam should follow the label encoding used during training, so the exact class meaning should be stated according to the dataset's LabelEncoder mapping.

Table 1: Comparison Result

	Accuracy	**Precision**	**Recall**	**F1-score**
Previous LSTM Technique	95%	94%	93%	-
Proposed Technique	96.93%	96.14%	90.88%	93.43%

Table 1 comparison shows that the proposed BERT + XGBoost technique outperforms the previous LSTM-based technique in terms of overall accuracy, precision, and F1-score. The proposed approach improves accuracy from 95% to 96.93% and precision from 94% to 96.14%. Although the recall of the proposed model is 90.88%, compared with 93% for the LSTM technique, the proposed method achieves a strong F1-score of 93.43%, demonstrating a good balance between precision and recall. Overall, the results indicate that BERT's contextual representation with XGBoost classification provides improved overall predictive performance compared with the previous LSTM approach.

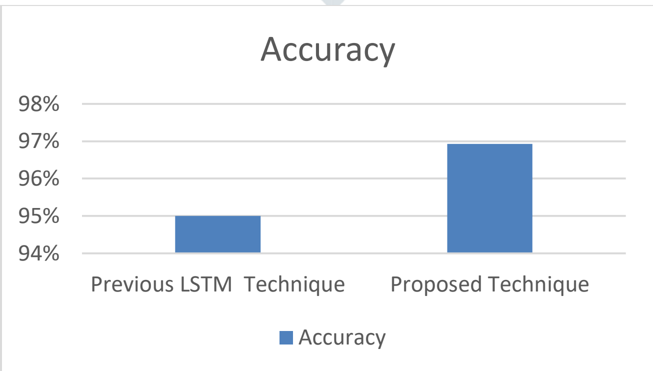

Figure 8: Accuracy

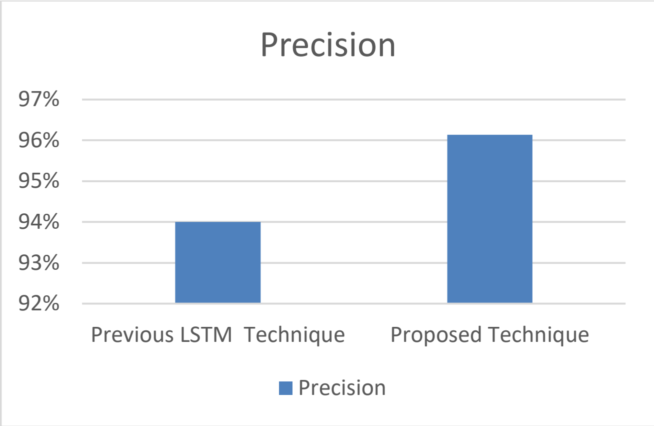

Figure 9: Precision

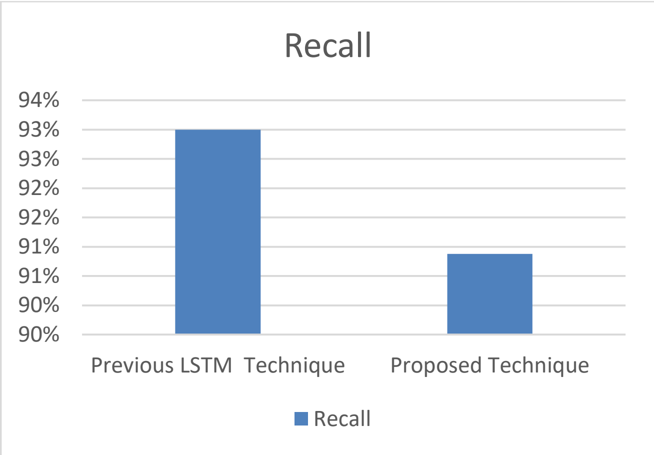

Figure 10: Recall

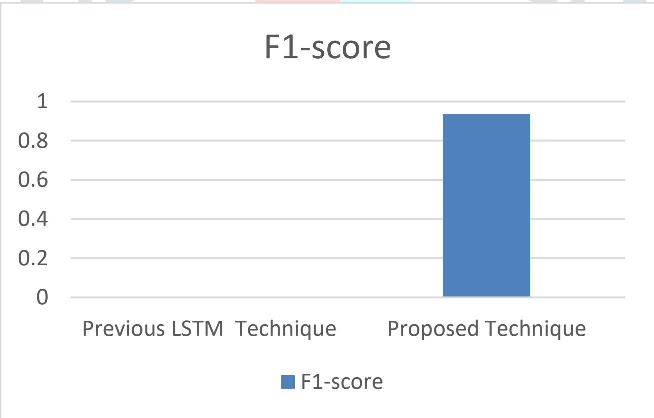

Figure 11: F1-score

V. CONCLUSION

This study presented an optimized BERT-XGBoost based framework for spam detection to improve the accurate identification and classification of relevant text spams from unstructured textual data. The proposed methodology combines the contextual feature extraction capability of BERT with the robust classification and optimization capability of XGBoost. BERT effectively captures semantic and contextual relationships within the text input, while XGBoost utilizes these representations to perform reliable spam classification. The proposed model achieved an accuracy of 96.93%, precision of 96.14%, recall of 90.88%, and F1-score of 93.43%. Moreover, the ROC-AUC of 99.50% demonstrates excellent discriminative performance. These results confirm that the hybrid BERT-XGBoost approach can effectively detect meaningful spams and provide a reliable solution for NLP-based information extraction. The proposed framework can be further extended to domain-specific applications such as healthcare, question answering, entity recognition, sentiment analysis, and automated information retrieval. Future work can focus on larger and more diverse datasets, hyperparameter optimization, multilingual spam detection, and comparison with advanced transformer-based architectures to further improve generalization and detection performance.

REFERENCES

[1] Muhammad Arif Syarifuddin Bin Ahmad, Muhammad Izzan Bin Rozlan, Athmar Danial Bin Yusri, “Spam Detection using Machine Learning Techniques”, *TechRxiv*. 12 February 2025.

[2] A. A. John, D. Olaniyan, J. Olaniyan, O. P. Bernard, M. Odighi and E. B. Michael, "Detecting Spam Emails Using Hybrid Machine Learning and Deep Learning Techniques," *2025 IEEE Connecting the Unconnected Europe, Middle East, and Africa Regional Summit (CTUS-EMEA)*, Abuja, Nigeria, 2025, pp. 1-6.

[3] G. Kanaan, M. Zghoul, G. Kanaan, A. M. Al-Zoubi and T. Kanaan, "Arabic Email Spam Detection Using Machine Learning and Deep Learning: A Comparative Study," *2025 12th International Conference on Information Technology (ICIT)*, Amman, Jordan, 2025, pp. 353-357.

[4] A. Yogaraj, T. R. D. Kumar, S. S. Babu, D. Hemanth Raj, A. Herrick Linore and P. Jagath, "Email Spam Detection using Machine Learning," *2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT)*, Kollam, India, 2025, pp. 904-909.

[5] E. H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi, and M. Uddin, “Email spam: A comprehensive review of optimized detection methods, challenges, and open research problems,” *IEEE Access*, 2024.

[6] P. L. R. Kameswari and P. R. Sudha Rani, "Multilingual Spam Classification Using Advanced Deep Learning Techniques," *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, Theni, India, 2024, pp. 1595-1601.

[7] M. Salman, “Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models,” in *IEEE Access*, vol. 12, pp. 24306–24324, 2024.

[8] K. Aparna, “Detection of Multilingual Spam SMS Using NaiveBayes Classifier,” *2023 IEEE 5th International Conference on Cybernetics*, Cognition and Machine Learning Applications, Hamburg, Germany, 2023.

[9] K. Debnath and N. Kar, "Email Spam Detection using Deep Learning Approach," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, 2022, pp. 37-41.

[10] H. Sharif, F. Rehman, and A. Rida, “Deep Learning: Convolutional Neural Networks for Medical Image Analysis-A Quick Review,” in 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), 2022, pp. 1–4.

[11] M. S. Ashraf, F. Rehman, H. Sharif, M. Aqeel, M. Arslan and A. Rida, "Spam Consumer’s Reviews Detection for E-Commerce Website using Linguistic Approach in Deep Learning," *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, Karachi, Pakistan, 2022, pp. 1-7

[12] S. Gadde, A. Lakshmanarao and S. Satyanarayana, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2021, pp. 358-362.

[13] G. Bathla and A. Kumar, "Opinion spam detection using Deep Learning," *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2021, pp. 1160-1164.

[14] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, “A comprehensive survey for intelligent spam email detection,” *IEEE Access*, vol. 7, pp. 168261–168295, 2019.

[15] G. M. Shahariar, S. Biswas, F. Omar, F. M. Shah, and S. B. Hassan, “Spam review detection using deep learning,” in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2019, pp. 27–33.