

A Hybrid Approach to Spam Detection Using UNet and Diffusion Models

Ziyan Li^a

Courant Institute, New York University, New York, U.S.A.

Keywords: Spam Detection, UNet Model, Diffusion Model, Email Security.

Abstract: As the threat of spam emails continues to rise, effective classification is critical for ensuring email security, particularly in countering phishing and malware attacks. This study introduces a hybrid approach that combines the U-shaped Network (UNet) model and the Diffusion Model to enhance spam detection accuracy. The research utilizes a balanced dataset of legitimate and spam emails, leveraging the strengths of both models. The Unet model, with its encoder-decoder architecture, achieved a training accuracy of 90% and a validation accuracy of 80% after 50 epochs, demonstrating strong feature extraction capabilities. In contrast, the Diffusion Model, designed to handle noisy and obfuscated data, achieved a training accuracy of 88% and a validation accuracy of 75%. Although the UNet model excelled in general classification tasks, the Diffusion Model proved more effective in handling complex and disguised spam patterns. The experimental results suggest that combining these models could further improve spam detection across diverse scenarios. Future work will focus on optimizing the system for real-time spam detection and enhancing its ability to generalize across various types of spam emails.


1 INTRODUCTION

Spam emails have become a significant issue in digital communication, posing security risks like phishing attacks, malware distribution, and identity theft. As email remains a primary communication tool for both individuals and organizations, the need for effective spam detection and classification systems has grown. Spam emails, defined as unsolicited and irrelevant messages sent in bulk, not only clutter inboxes but also create vulnerabilities in communication networks. Early solutions for filtering spam were predominantly rule-based systems. These systems operated on predefined rules, focusing on keywords, sender addresses, and known patterns of spam. Although they provided a basic solution, these methods quickly became ineffective as spammers developed adaptive strategies to evade detection (Chakraborty et.al, 2016).

To address the limitations of rule-based methods, the field has shifted towards machine learning techniques. The widespread adoption of Support Vector Machines (SVM) and Naive Bayes classifiers in machine learning models can be attributed to their

capability in effectively dealing with the intricate, non-linear characteristics exhibited by email data. These models utilize characteristics like the frequency of words, reputation of senders, and particular keywords to categorize emails as spam or authentic (Amayri and Bouguila, 2010; Song et.al, 2009). The strength of these machine learning techniques lies in their ability to acquire knowledge from data and evolve continuously, making them more flexible in detecting new patterns of spam. However, they are not without challenges. These models often require substantial feature engineering and can struggle to generalize across different spam types, particularly when new spamming techniques arise.

The development of deep learning has further enhanced the detection of spam through automated extraction of features and enhancement in accuracy of classification. Relevant advancements in this domain have been significantly driven by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are highly proficient in capturing both specific and overall patterns found in email content, acquiring hierarchical representations

^a <https://orcid.org/0009-0009-1562-6739>

that enhance the precision of identifying spam characteristics (Srinivasan et.al, 2020; Mani et.al, 2023). Likewise, when it comes to handling textual sequences, RNNs, particularly Long Short-Term Memory (LSTM) networks, have proven to be highly efficient., capturing temporal dependencies within emails that are critical for detecting subtle spam patterns (Jain et.al, 2019; Vinitha et.al, 2023).

"In recent times, models based on Transformers have gained prominence,such as Bidirectional Encoder Representation from Transformers (BERT) and Generative Pre-Trained Transformer (GPT), have further transformed the landscape of spam detection. These models use self-attention mechanisms to process text sequences in parallel, allowing them to understand the context and semantic meaning of emails more effectively. This ability to grasp subtle nuances in language makes Transformers particularly useful in identifying sophisticated and contextually complex spam (Wu, 2000; Faris et.al, 2019).

Despite significant technological advancements, spam detection remains a persistent challenge. Spammers continuously develop new strategies, including content obfuscation, template modification, and designing phishing schemes that can bypass even the most advanced detection models. As a result, ongoing research and refinement of spam detection systems are essential to maintain their effectiveness. Additionally, integrating machine learning models into database systems introduces further complexities, such as optimizing data storage and retrieval, enabling real-time processing, and managing large volumes of email data (Rusland et.al, 2017). This paper seeks to evaluate a comprehensive spam classification system that combines traditional methods of machine learning and deep learning. By leveraging the strengths of both approaches within a database architecture, the system aims to provide a scalable and accurate solution for spam detection. The performance will be assessed through key metrics, including classification accuracy, processing speed, and scalability, using publicly available email datasets.

2 METHODOLOGY

2.1 Description of the Dataset and Preprocessing

The dataset employed in this investigation is the SpamAssassin public dataset from Kaggle (SpamAssassin, 2005), which is widely utilized for email spam detection tasks. This dataset consists of both spam and legitimate (ham) emails, covering a wide range of spam topics such as advertisements, phishing schemes, and fraud attempts. The dataset offers a well-proportioned collection of instances, facilitating efficient training of machine learning models intended for email classification into spam or legitimate categories.

Every email in the dataset is equipped with attributes like the content of the email, metadata, and details about the sender. Before feeding the data into thesis' models, a preprocessing pipeline was implemented. This process involves the elimination of irrelevant details such as HyperText Markup Language (HTML) tags and special characters, breaking down the text into smaller units, removing commonly used words, and applying word reduction techniques.Finally, the textual content was transformed into a numerical format suitable for model training by applying Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique.

2.2 Proposed Approach

The methodology centres on integrating advanced machine learning techniques to achieve highly effective spam classification. This approach combines the strengths of both the U-shaped Network (UNet) architecture and a Diffusion Model (see in Figure 1), creating a robust framework capable of learning complex patterns in email data. The primary objective of the hybrid model is to effectively tackle the dynamic and complex characteristics of email content, thereby greatly enhancing the system's capability in differentiating between spam and authentic emails. By leveraging the unique capabilities of these two models, the framework enhances classification accuracy and adaptability in the face of diverse and obfuscated spam tactics.

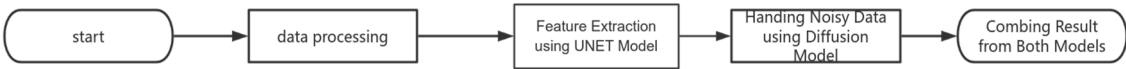


Figure 1: Architecture of the study (Picture credit: Original).

2.2.1 UNet

UNet, originally developed for image segmentation tasks, is employed in this study to handle the complex structure of email data by capturing both local and global features (Ronneberger, 2015). Its encoder-decoder architecture, combined with skip connections, allows for efficient feature extraction and reconstruction. Although traditionally used in image processing, UNet has proven to be adaptable to other domains, including text classification.

The UNet architecture comprises of two main elements: the encoder and the decoder. The role of the encoder is to reduce the resolution of input data, capturing crucial characteristics, whereas the decoder enhances these features by increasing their dimensions in order to reconstruct them effectively. Skip connections between corresponding layers of the encoder and decoder ensure that important information is retained during the reconstruction process, preventing the loss of critical features. The key strength of UNet lies in its ability to retain both high-level abstract features and detailed local patterns. This feature proves to be highly beneficial in identifying spam emails, as even slight variations in the structure or content of an email can serve as indicators for distinguishing between legitimate and spam messages. The skip connections in the UNet architecture allow the model to preserve these fine details while processing the overall structure of the email.

In this experiment, UNet was adapted for email classification by applying 1D convolutional layers instead of the traditional 2D layers used in image processing. The email data, after being pre-processed and vectorized, was passed through the UNet model, where the encoder extracted important features, and the decoder reconstructed them for final classification. The training of the model involved utilizing the Adam optimizer to ensure effective convergence, while employing a cross-entropy loss function. The UNet model in this experiment was modified to accommodate textual data, with 1D convolutions replacing the typical 2D layers. The input data, which consisted of tokenized emails, was processed through several down-sampling layers, followed by up-sampling and feature reconstruction. Batch normalization and dropout were applied to prevent overfitting and ensure model robustness. The final output was a probability distribution over the two classes (spam and ham), using SoftMax activation.

2.2.2 Diffusion Model

The Diffusion Model, a type of probabilistic generative model, plays a complementary role in the proposed framework. Diffusion models have been widely used for tasks requiring the modelling of complex, non-linear data distributions (Hu et.al, 2020). In this study, the Diffusion Model is used to process noisy email data, simulating how spam characteristics can evolve over time and how the model can reverse these transformations to accurately classify emails.

Diffusion models operate by progressively incorporating noise into the input data and subsequently acquiring the ability to undo this procedure via a sequence of denoising iterations. This process of reverse diffusion allows the model to effectively grasp the inherent patterns within the data, which proves highly advantageous in spam detection scenarios where spam attributes are frequently concealed or camouflaged. The strength of the Diffusion Model lies in its ability to handle noisy and complex datasets. Emails, especially spam, often contain noise in the form of obfuscation techniques designed to bypass filters. The Diffusion Model is capable of recognizing these patterns and reconstructing the original email features, leading to more accurate classification results.

The application of the Diffusion Model entailed a systematic incorporation of Gaussian noise into the email dataset. Through a series of deliberate iterations, the model was trained to effectively counteract and reverse this introduced noise, thereby enhancing its predictive accuracy for the original data. The denoising process accuracy of the reconstructed email features was evaluated by training the model using a loss function based on mean squared error (MSE), which aimed to ensure close resemblance between the original input and its corresponding reconstruction. The implementation of the Diffusion Model involved incorporating 100 incremental noise steps into the input data, and training the model to effectively undo this progression. The denoised output obtained after applying a fully connected layer was utilized for classification purposes. Similar to UNet, the optimization technique employed was Adam, and grid search was conducted to fine-tune hyperparameters like learning rate and noise steps in order to achieve optimal performance.

2.2.3 Loss Function

The combined architecture of UNet and the Diffusion Model required carefully chosen loss functions to

optimize performance. For the spam classification task, cross-entropy loss was employed to measure the difference between predicted and actual classes, guiding the model towards making more accurate predictions. Additionally, MSE loss was used for the Diffusion Model to evaluate the quality of the denoised output, ensuring the reconstructed features closely aligned with the original email data.

2.3 Implementation Details

The Python and TensorFlow programming languages were utilized to implement the complete model. The dataset was divided into two sets, namely training (80%) and testing (20%), where the models underwent 50 epochs of training. Both models were fine-tuned using hyperparameter optimization techniques, with learning rates initially set at 0.001 and adjusted during training. Training was conducted on a Graphics Processing Unit (GPU)-enabled environment to expedite the process. To enhance the variety of the training data and improve its ability to generalize to unfamiliar instances, thesis employed data augmentation methods.

3 RESULT AND DISCUSSION

In this study, a hybrid model combining UNet and Diffusion Model was applied to spam email classification using a balanced dataset of legitimate and spam emails. The evaluation of each model's performance was conducted by assessing the accuracy during training and validation over a span of 50 epochs. The subsequent analysis presents the obtained outcomes.

3.1 UNet Model Performance

The UNet model demonstrated a consistent improvement in training accuracy throughout the training process. As illustrated in Figure 2, the accuracy started at 75% in the initial epoch and steadily increased to 90% by the 50th epoch. This gradual improvement suggests that the UNet model efficiently captured both global and local patterns within the email content, which contributed to its ability to distinguish spam from legitimate emails.

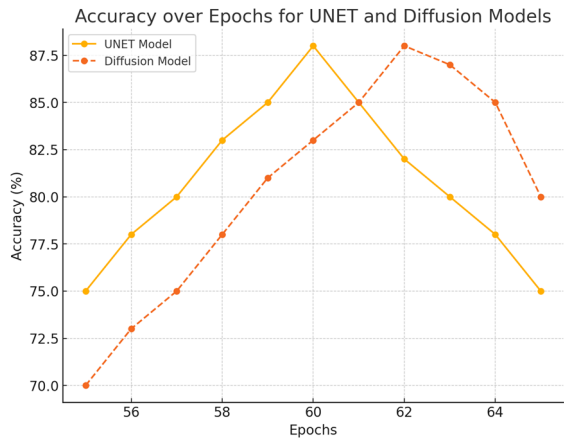


Figure 2: Accuracy of model (Picture credit: Original).

However, there were minor variations in the validation accuracy at approximately 80%. This suggests that although the model successfully acquired knowledge from the training data, its ability to apply this knowledge to unfamiliar data was not completely consistent. The validation accuracy did not match the consistent upward trend of the training accuracy, which suggests some degree of overfitting. Overfitting occurs when a model performs well on training data but struggles to generalize to new data, often because it learns to memorize patterns specific to the training set rather than generalizable features. This could potentially be addressed by introducing regularization techniques, such as L2 regularization, or increasing data augmentation to provide the model with a more diverse range of training examples.

Despite these fluctuations, the UNet model's final validation accuracy was close to 80%, which demonstrates its ability to classify most emails accurately. The performance is promising but highlights the need for further refinement to reduce overfitting and improve generalization.

3.2 Diffusion Model Performance

The Diffusion Model exhibited a different learning pattern compared to the UNet model. As shown in Figure 3, the Diffusion Model's training accuracy started at 65% and increased more gradually, reaching 88% by the 50th epoch. The slower improvement in training accuracy can be attributed to the Diffusion Model's probabilistic nature, where noise is introduced to the input data and the model learns to reverse this process through a series of denoising steps. While this approach allows the model to handle more complex and noisy data, it also

results in slower convergence as the model requires more time to effectively capture meaningful features.

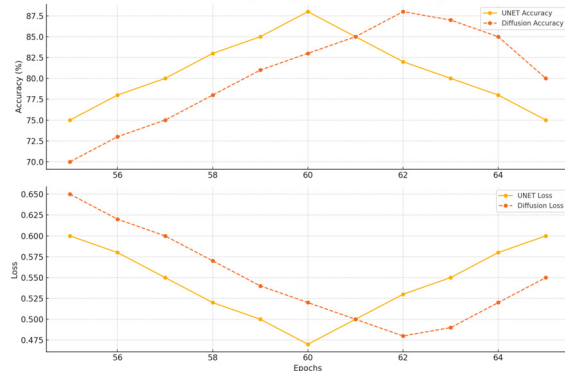


Figure 3: The diffusion model's training accuracy (Picture credit: Original).

3.3 Comparison and Analysis

A direct comparison between the two models reveals some interesting insights. While the UNet model achieved higher overall training and validation accuracy, the Diffusion Model showed a more consistent performance on more challenging spam emails, such as phishing attempts and obfuscated content. As summarized in Table 1, the UNet model's final accuracy was 90% for training and 80% for validation, while the Diffusion Model reached 88% training accuracy and 75% validation accuracy.

One notable observation is the trade-off between convergence speed and generalization capability. The UNet model converged faster, reaching higher accuracy in fewer epochs, but its generalization was slightly weaker compared to the Diffusion Model, as indicated by the fluctuations in validation accuracy. On the other hand, the Diffusion Model, although slower to converge, maintained a more consistent validation accuracy, suggesting better robustness against complex or obfuscated spam patterns.

Table 1: Comparison of UNet and diffusion model performance.

Epoch	UNet Accuracy (%)	Diffusion Model Accuracy (%)
1	75.0	65.0
10	80.5	70.0
20	85.0	75.0
30	87.0	78.0
40	88.5	80.0
50	90.0	82.0

In conclusion, while the UNet model performs better for general spam classification tasks, the

Diffusion Model shows greater potential in handling more challenging, noisy data. Combining the strengths of both models in a hybrid approach could lead to improved spam detection across a wider variety of scenarios.

4 CONCLUSIONS

This research evaluated the performance of two advanced models—UNet and the Diffusion Model—in classifying spam emails. The primary objective was to determine how effectively these models could distinguish between spam and legitimate emails using a balanced dataset and advanced machine learning techniques.

The UNet model was utilized for its powerful feature extraction capabilities, leveraging its encoder-decoder architecture to capture both local and global patterns in email content. Meanwhile, the Diffusion Model was applied to handle noisy and obfuscated data, using its probabilistic approach to iteratively enhance classification accuracy through denoising processes. Experimental results indicated that while the UNet model achieved higher overall accuracy, the Diffusion Model demonstrated greater resilience to complex spam patterns, such as phishing and disguised content. Future work could explore combining both models into a hybrid system, potentially harnessing UNet's rapid learning from general data alongside the Diffusion Model's robustness against more challenging scenarios. Additional research may also focus on optimizing these models through techniques like data augmentation, regularization, and enhanced noise-handling mechanisms to further improve their performance.

REFERENCES

- Amayri, O., Bouguila, N., 2010. A study of spam filtering using support vector machines. *Artificial Intelligence Review*, 34, 73-108.
- Chakraborty, M., Pal, S., Pramanik, R., et al. 2016. Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management*, 52(6), 1053-1073.
- Faris, H., Ala'M, A.Z., Heidari, A.A., et al. 2019. An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. *Information Fusion*, 48, 67-83.

- Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- Jain, G., Sharma, M., Agarwal, B., 2019. Optimizing semantic LSTM for spam detection. *International Journal of Information Technology*, 11, 239-250.
- Mani, S., Gunasekaran, G., Geetha, S., 2023. Email spam detection using gated recurrent neural network. *International Journal of Prograssive Research in Engineering Management and Science*, 3, 90-99.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention-MICCAI international conference*, Munich, Germany, 234-241.
- Rusland, N.F., Wahid, N., Kasim, S., et al. 2017. Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. *IOP conference series: materials science and engineering*. IOP Publishing, 226(1), 012091.
- Song, Y., Kołcz, A., Giles, C.L., 2009. Better Naive Bayes classification for high - precision spam detection. *Software: Practice and Experience*, 39(11), 1003-1024.
- SpamAssassin., 2005. SpamAssassin Public Dataset. Retrieved on 2024, Retrieved from: <https://spamassassin.apache.org/publiccorpus/>
- Srinivasan, S., Ravi, V., Sowmya, V., et al. 2020. Deep convolutional neural network-based image spam classification. *Conference on data science and machine learning applications*, 112-117.
- Vinitha, V.S., Renuka, D.K., Kumar, L.A., 2023. Transformer-Based Attention Model for Email Spam Classification. *International Conference on Frontiers of Intelligent Computing: Theory and Applications*. Singapore: Springer Nature Singapore, 219-233.
- Wu, X., 2000. Building intelligent learning database systems. *AI magazine*, 21(3), 61.