ISSN: 2093-5374 / E-ISSN: 2093-5382

Artificial Intelligence-Based Multi-Tiered Architecture for the Detection of Fake News, Spam Data and Unauthorized Users

P. Kardeepa[1], N. Subbulakshmi[2*], and A.M. Gurusigaamani[3]

[1]Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India. p.kardeepa@klu.ac.in, https://orcid.org/0000-0002-6354-8175

[2*]Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India. subbulakshminammalwar@gmail.com, https://orcid.org/0000-0002-9863-5033

[3]Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India. gurusigaamani@klu.ac.in, https://orcid.org/0000-0001-8652-7468

Received: October 21, 2025; Revised: December 19, 2025; Accepted: January 30, 2026; Published: March 31, 2026

Abstract

The growth of online platforms has triggered the growing necessity of effective detection systems to deal with threats such as fake news, spam, and unauthorized users. Existing models have serious problems, such as limited task coverage, limited generalization across diverse datasets, overfitting, and the inability to address multi-domain threats within a single framework. As a way to overcome these obstacles, we would suggest the Multi-Tiered Architecture for Spam Data Detection (MTA-SD), a solution that will incorporate AI-based methods, including machine learning, deep learning, and natural language processing, into a multi-layered detection architecture. The system is able to address more than fake news, spam, and unauthorized user detection, but by augmenting digital platforms' security, it supports the objective of SDG 9: Industry, Innovation, and Infrastructure, by providing safe, robust, and innovative infrastructures. The architecture addresses three major detection tasks simultaneously, including Fake News Detection (FND), Spam or Malicious Data Detection (MDD), and Unauthorized User Detection (UUD). Using the hybrid systems of BERT with textual data, GPT-2 with contextual insights, and XGBoost for spam classification, the proposed system is guaranteed to achieve high accuracy across different types of inputs. The results of the evaluation indicate that the MTA-SD model is better than the previous solutions since it has an accuracy of 99.90, precision of 99.95, recall of 99.96, and an F1 score of 99.99% in various datasets, including ISOT, LIAR, IFND, and Malicious Webpages. This capability of the architecture to combine various detection activities, make use of sophisticated feature engineering, and evolve with new threats with the continuous learning of new features makes it a scalable and robust system to achieve real-time detection. This model not only addresses the constraints of the traditional methodologies but also offers a multifaceted, flexible, and very precise system for solving the current issues of digital security.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March-2026), pp. 743-765. DOI: *10.58346/JOWUA.2026.I1.041*

*Corresponding author: Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Virudhunagar, India.

Keywords: Artificial Intelligence, Multi-Tiered Architecture, Fake News Detection, Spam Detection, Unauthorized User Detection, Machine Learning, Deep Learning, Natural Language Processing, BERT, GPT-2, XGBoost, Autoencoder, Data Security, MTA-SD, Digital Innovation.

1 Introduction

Today, the ability to produce large amounts of data through digital platforms and social media is increasing rapidly. With this increase comes the risk of false information and spam, as well as the fear of unverified users accessing that data (Al-Haija & Droos, 2025; Hossen et al., 2025). To combat these issues, we have developed an artificial intelligence-based multi-tiered architecture that utilizes machine learning and deep learning to identify and remove harmful content across multiple platforms. The multi-tiered design's main function is to develop an automated system to identify and categorize fake news, spam messages, and activity by unauthorized users fast (Chandrika & Raju, 2025; Mounika & Reddy, 2025; Khalil et al., 2017). The important aspect of the architecture is that it consists of multiple layers, with each layer having its own distinct role. The first tier focuses on gathering and processing data from many different sources, while each level after uses artificial intelligence-based algorithms to detect and flag any issues or harmful content. To determine whether news is "fake," we use both natural language processing and machine learning to analyze each source's context, sentiment, and reliability (Rasul & Jumaa, 2022; Szczepański et al., 2021; Jwa et al., 2019; Soy & Balkrishna, 2024). The same methods are used to detect spam; however, we primarily use pattern recognition to locate unwanted and/or harmful content. While Unauthorized User Detection (UUD) involves analysing user activity and verifying identity to identify possible threats resulting from unauthorized access to the system, it does so in a manner that allows these components to be integrated into a cohesive unit or solution (Qin & Zhang, 2024; Farokhian et al., 2024). The tiered approach to building this solution is predominantly AI-driven and quite versatile/scalable in nature, providing capability for working with high volumes of data as seen in current digital environments, where maintaining quality data and protecting users is paramount (Al Ghamdi et al., 2024). In addition to further enhancing the performance of UUD solutions, this architecture also helps prevent problems such as false or misleading information, data misuse, and unauthorized entry (Lin, 2024).

The possibility to create and share large quantities of data via digital platforms has led to great advantages, yet to the increased threat of fake news, spam, and unauthorized access. In order to overcome these difficulties, it is important to develop sound detection systems to ensure the integrity and safety of the digital infrastructures. This study corresponds to SDG 9: Industry, Innovation, and Infrastructure, which stresses the need to promote innovation, enhance digital infrastructure, and have resilient and secure systems. This work will contribute to the purpose of creating resilient infrastructure and advancing innovation in the field of digital security, thereby making digital platforms safer and more reliable through an AI-based, multi-tiered detection architecture to ensure the safety of the digital infrastructure.

Key Contribution

- **Multi-Tiered Detection System:** The system's structure consists of multiple tiers that are each designed to address a specific category of threat (evolved from traditional sources of misinformation to bots/automated accounts/bots ending up in the arena of cybersecurity). The multi-tiered structure enables the system to provide more precise detection, thereby strengthening its ability to identify fraudulent and malicious activity.

- **AI-Driven Detection:** The combination of machine learning, deep learning, and natural language processing (NLP), using AI mechanisms, provides enhanced capability to detect threats and adapt to ongoing changes in user behaviour by learning from data patterns that evolve over time.
- **Real-Time Detection:** The ability of the system to detect and prevent illicit content, and potentially spam, as well as user activity, is critical to mitigating the proliferation of misinformation/fake news and bad/negative behaviours in real-time.
- **Scalable and Adaptive:** The architecture is designed to be optimally managed to handle a large volume of information and also allows for the modification of characteristics of the architecture to support emerging trends and evolving characteristics of its user base. Companies that have platforms that see the growth of numerous new users and a range of user behaviour and preferences will benefit significantly from building their solutions on this architecture.
- **Comprehensive Security and Data Integrity:** By combining multiple types of problems in a single platform, this solution dramatically enhances the level of security and data integrity that its clients will experience, as well as increasing the level of confidence that will result from using this solution.

This research paper comprises several sections. Each section should have various concepts. Section I introduces the research topic, followed by the detection of fake news, spam data, and unauthorized users. This section also includes the key contribution, followed by the main objective of this research. Section II describes the Literature review section, followed by the previous papers. Section III describes the Proposed Methodology, which consists of an overall architecture diagram, explains the working principle for integrating with GPT and BERT for fake news detection, Generative pre-trained transformer Components, XGBoost for MDD, and an Auto Encoder for UUD. In this, the proposed algorithm is also explained. Section IV explained that the Results and Discussion section consists of hardware and software configuration, Performance Evaluation, followed by dataset description, Various dataset sample Distributions, and performance comparison of various models using datasets LIAR, ISOT, IFND, and Malicious Website Dataset. Section V describes the conclusion of the main research findings.

2 Literature Review

The emergence of misinformation, fake news, spam data, and unauthorized users on social media and digital platforms in recent years has prompted the need to use automated detection systems more than in the past (Saminathan et al., 2023; Dubcek & Voronina, 2021). The purpose of these systems is to ensure that the users of the system are not misled, maintain trustworthiness, and reduce the security risks (Takiddin et al., 2022; Zheng et al., 2021). Multi-tiered architectures have been used with great success, and one of the technologies that has become useful in this direction is Artificial Intelligence (AI). Fake news, spam, and unauthorized user detection systems based on AI tend to employ a variety of methods. The content-based detection layer is concerned with text and multimedia content analysis in case of manipulation or falsification (Al-Rawi et al., 2019; Mokoena & Nilsson, 2023). Most of the machine learning classifiers, such as Naive Bayes, Support Vector Machines (SVM), and Random Forests, have typically been applied to feature extraction, which can consist of bag-of-words models, n-grams, or metadata-based features. Nevertheless, further sophisticated deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their hybrids, have been utilized due to their capability to learn complex patterns and latent semantic associations in data (Park & Choi, 2025). Also, multimodal strategies, the analysis of both textual and image data, have

become prominent, especially in the detection of misinformation that is related to images (Babu et al., 2023; Balakrishnan & Leema, 2025). Transformer-based models such as BERT have also been broadly embraced due to their better language semantics and context management, and even the plausibility of the sources, which is important in detecting fake news. Such characteristics as the age of the accounts, the frequency of the posts, and unusual posting patterns are examined to identify the genuine customers and possible spammers or bots. Certain systems are also concerned with metadata of URLs, which may be reflective of phishing or spam URLs (Sharma et al., 2025; Khule et al., 2025). The other important tier is network and social context analysis, in which fake news or spam are likely to be shared in strange or organized forms across social networks. In this case, AI systems are based on graph models, which examine the connection between users and their activity on social media and isolate outliers and suspicious interactions. These features on the network level are essential when it comes to the identification of bot networks or organized misinformation (Lim et al., 2025). This mixture of these methods creates hybrid and multi-level models that give a better detection mechanism. These models can identify online deception of different types effectively by combining content analysis, user behavior, and network analysis. As an example, other systems integrate classical machine learning with deep learning, network analysis, and user metadata, and can more accurately identify less noticeable examples of fake news or spam (Udayakumar et al., 2023). Moreover, there are systems that use weakly supervised, semi-supervised, or online learning methods due to the problems of changing modes of deception and scanty labelled data. The models are more resilient to new forms of attacks because they are able to continually adapt using the feedback or reports provided by the users or based on the new patterns (Naghib et al., 2025). The multi-tiered architecture approach is an effective solution to all these problems as it is based on overlaying the various types of analysis that include: content, user behavior, and network propagation in a single framework (Christy et al., 2025). This type of system is effective in increasing the identification of fake news, spam, and unauthorized users, in addition to expanding the scalability and flexibility of the solution. With the continuous multimodality and context sensitivity of fake news and spam, it is likely that in the future, the multimodal AI systems will evolve to support more text, images, videos, and metadata. Moreover, it is becoming increasingly important to make AI-based detection systems explainable because such decisions as banning accounts or flagging content may lead to serious consequences. Therefore, it will be important to improve the clarity and explainability of AI models to enable trust and responsibility in automated detection systems. To sum up, AI-powered multi-tiered architectures are a potent way to address the current multiplying issues of fake news, spam, and unauthorized users. These systems can be used to identify deceptive content over a wide range of platforms with scalable, adaptable, and robust solutions that are made possible by a combination of machine learning, deep learning, and network analysis algorithms (Someswara Rao et al., 2024; Yi et al., 2025). Nevertheless, issues like data quality, model explainability, and adversarial behavior should be brought to the table so that these systems can be efficient regardless of the changing threats.

Research Gap

One of the research gaps that has been detected in the literature is that it has achieved a lot in the field of AI-based multi-tiered architectures in detecting fake news data, spam content, and unauthorized users; there are still some gaps. Firstly, more robust data sets should be developed that are heterogeneous, representative, and unbiased to enhance better model generalization and performance in various fields. Second, deep learning models have been promising, but a critical problem is the explainability of these models, especially when the consequences of such decisions (e.g., a ban on the user or a content flag) are serious. Moreover, real-time flexibility of the detection systems to new strategies of attack, e.g., AI content, dynamic spam actions, is a field that still has to be explored further. The multimodal integration

(text, images, videos, and user data) requires further research because integration of these various sources of data may increase detection, yet adds complexity to model design and evaluation. These weaknesses bring a new avenue of development on the detection systems and their use in practice.

3 Proposed Methodology

3.1 Overall Diagram for Proposed Methodology

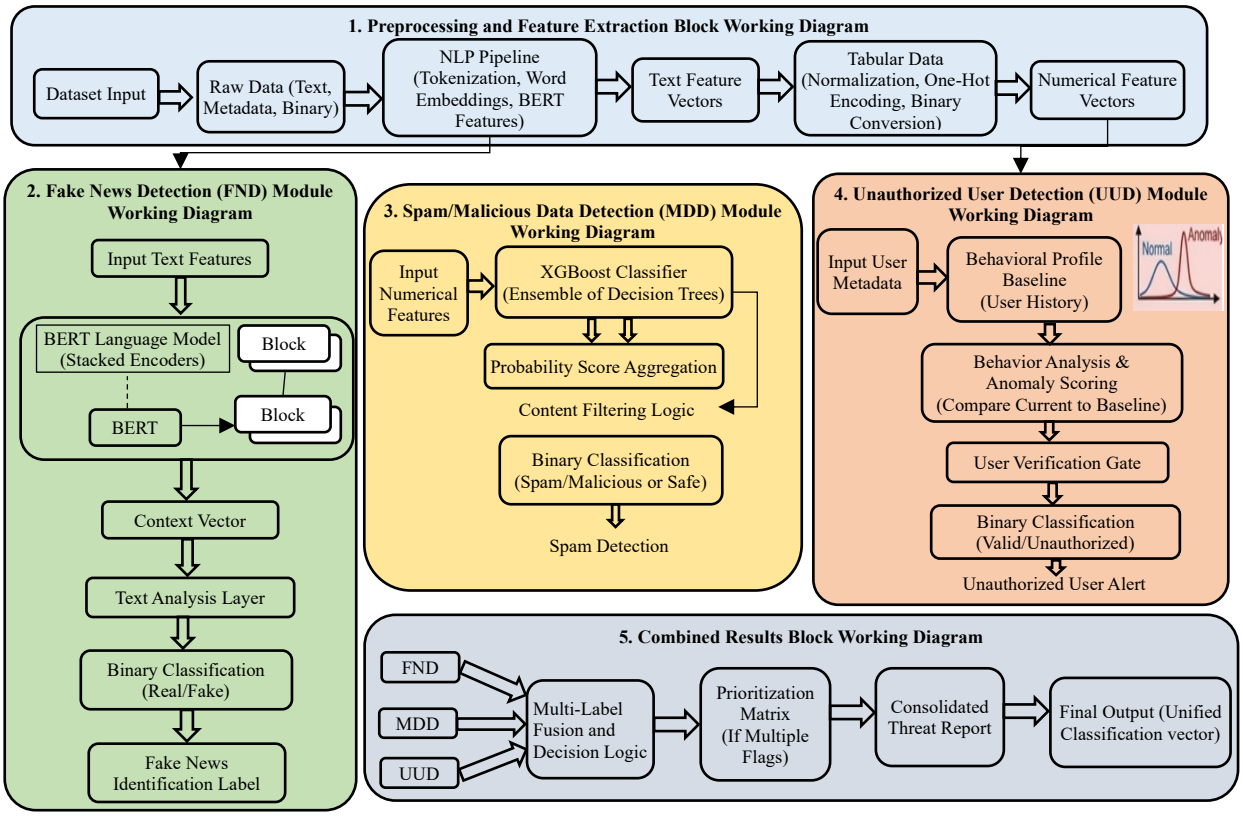

Figure 1: Block diagram for proposed methodology

The diagram (Figure 1) shows how the proposed model would work as a whole for different detecting tasks. The "Dataset Input" is at the middle, and it is the model's main input. The model then splits into three main detection tasks: Fake News Detection (FND), Spam or Malicious Data Detection (MDD), and Unauthorized User Detection (UUD). The tasks are shown as separate modules that work at the same time, each one focusing on a different form of detection. The results of these detection tasks are then put together to make the "Output" step. This architecture provides a multi-task model that can solve many detection problems at the same time. For example, it can find fake news, spam or malicious data, and unauthorized users, and then give a complete output based on all of these jobs. This method shows how flexible and adaptable the proposed paradigm is when it comes to dealing with a wide range of data security and integrity issues.

3.1.1 Integrating GPT and BERT for Fake News Detection

This architecture (Figure 2) provides the clear concepts of methodology with contains the fake news identification, which is characterized by the supervised followed by to determine particular items was fake or authentic.

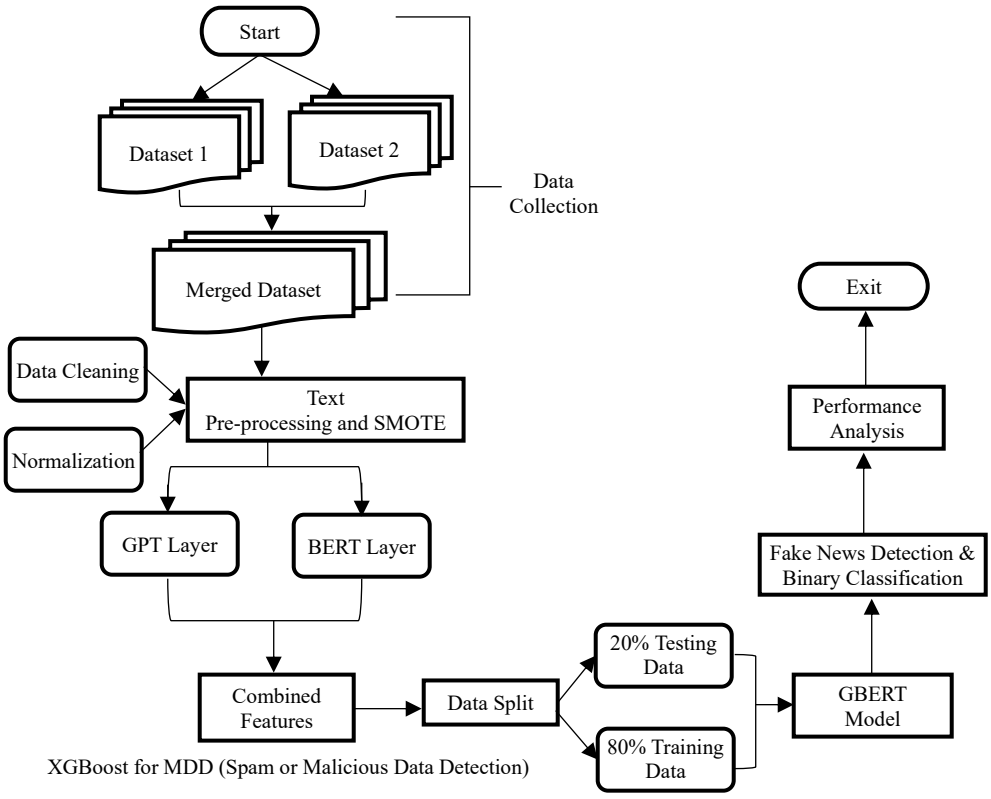

Figure 2: Integrating GPT and BERT for fake news detection

$$Let\ T = \{A1, A2, A3, \ldots\ldots\ldots An\} \quad (1)$$

The above equation (1) represents the n news articles that comprise only the textual information, followed by the various textual information with $L \in \{0,1\}$. This also represents the new label: 0 should be noted as real, and 1 as fake. The aim of the prediction model Function should be expressed as a feature vector $X(A)$, which contains the input to predict the label. $F(A) \rightarrow \{0,1\}$. From the above $F(A) = 0$, noted A as predicted as real and $F(A) = 1$, which means noted as fake.

The rapid development of NLP and DL has generated a lot of interest. This report discussed a hybrid deep learning framework that combines the strengths of BERT and GPT to evaluate the efficiency of detecting fake news. It also has three components: the network layer, GPT, and BERT. The acronym BERT can be fully explained by its components: Bidirectionality means that, as a context model, it can read and understand the written text input in both LTR and RTL directions, and can do so simultaneously. Encoder Representations: As per the architecture shown in figure 2 above, the encoder contains several feed-forward and self-attention neural units arranged in layers. For the model to grasp the semantic relationship within a sentence, this is crucial. The BERT model is built on the Transformer architecture, and its ability to understand and manipulate natural language text so efficiently and accurately is what makes it a true transformer. The model can access and utilize various contextual information from the input text, can work effectively with sequences of varying length (of arbitrary length), and can form powerful representations of words via self-attention. To establish associations with the input and output components, attention is used.

3.1.2 Generative Pre-trained Transformer (GPT) Components

GPT is the most suitable model for NLP applications, it is the direct contextual link among the long-range dependencies. This also helps anticipate gaining syntax, semantics, and context from unsupervised, pretrained datasets.GPT can be trained on smaller labelled datasets to perform better on specific tasks due to transfer learning. The GPT-2 Transformer architecture, with 1.5 billion parameters, made a significant impact when it was released. In GPT architecture, which is comparable to the decoder component of transformer architecture (Figure 2), each token is output sequentially and added to the input sequence until the statement's end. Using masked self-attention, GPT can peek at future tokens as it processes each one to understand their link. Various iterations of the GPT-2 design exist, each with a unique dimensionality and number of decoders. Due to resource constraints, the GPT-2 model (small) is used in this work

3.1.3 XGBoost for MDD (Spam or Malicious Data Detection)

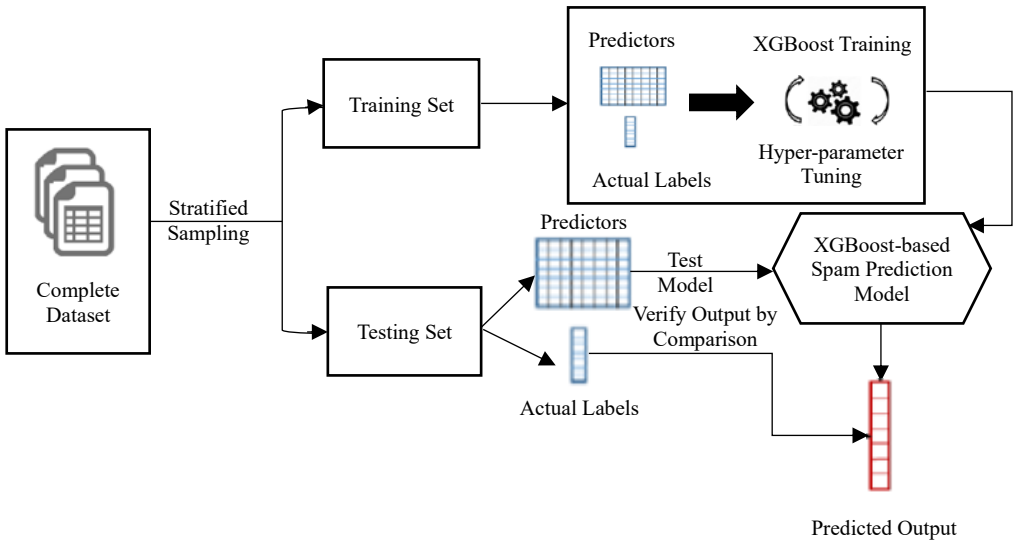

Figure 3: XGBoost for MDD (Spam or malicious data detection)

Figure 3 depicts the conceptual architecture of the experimental setup and the model implementation for the proposed spam detector using XGBoost. The obtained dataset was divided into a training set and an out-of-sample test set at a 7:3 ratio, using stratified sampling. The training set's 70% primary purpose is to familiarize the model with a wide range of spam and non-spam email types, while the testing set's 30% primary purpose is to ensure the proposed model is evaluated on data it has not previously encountered. The model was implemented in the R programming environment using the XGBoost package. The model was trained on the training set containing the real labels using leave-one-out cross-validation. This training method involves using only a portion of the training set to train the model, then validating it on the remaining data. The ability to assess the optimized model's performance during training makes this method crucial for hyperparameter tuning. Because of this, hyperparameter optimization was performed all at once during model training. Furthermore, determining the best hyperparameters for a machine learning model on a particular dataset can be difficult, and XGBoost is no exception. In fact, developing XGBoost models may make this work much more difficult due to the vast number of configurable hyperparameters available. Consequently, the spam detection model underwent a grid search over a predetermined set of hyperparameters. What follows is an algorithmically

highlighted version of the grid search technique: 1. Establish a collection of hyperparameters, such as eta, gamma, minimal kid weight, etc., and their potential values. 2. Go through each possible combination of hyperparameters sequentially. Go on to step 2 to record performance data; If there are more combinations of hyperparameters than that, go back to step 2. Train using the next set of hyperparameters in the sequence. 3. Determine the optimal combination of hyperparameters and construct the model. As with earlier work with XGBoost, we have limited the grid search to a subset of the hyperparameters for improving tree performance due to the large number of them and the time needed to optimize them all. When the variables of eta, gamma, maximum depth, and column sample were combined to the proposed spam detector performed at its best. We also set rounds to 200, which is the number of trees to grow; if the training error does not improve after 10 rounds, we quit early. The purpose of this regularization is to prevent overfitting of the training data. All three of the other booster parameters were set to 1, including the minimum sum of instance weight required in a kid and the proportion of data instances to be used while building trees.

The XGBoost approach is used for supervised learning in ML, which is an ensemble of weak classification and regression trees. CART is a unique type of decision trees that combines the real score among each leaf, the prediction score of each tree should be evaluated through K with the additive functions of f_k with the space for all possible functions of CART F as,

$$y^{\wedge} = \sum_{k=1}^{k} f_k(x_i), f_k \epsilon F \tag{2}$$

Equation (2) above describes the objective function, followed by the first and second terms based on the differentiable loss function. Which is the measure of the difference predicted by $y_i^{\wedge}$ target as y_i, which mentions the regularization term of the measure with the model complexity.

$$Obj\ (\theta) = \sum_{i}^{n} (y_i\ , y_i^{\wedge}) + \sum_{k}^{K} \Omega(f_k) \tag{3}$$

Equation (3) describes $\Omega(f) = \text{Y}T + \frac{1}{2}\lambda \sum_{j-1}^{T} w_j^2$ The vector score of each leaf should be represented by w and T, respectively. An additional thing, as constants of Y and λ Included with the control is the degree of regularizations. Our methods are designed to prevent overfitting in KGBoost, including feature and data subsampling. An additional point, regarding training these employees in XGBoost via prediction, should be formulated as equations 4 and 5.

$$y_i^{\wedge(t)} = \sum_{k=1}^{K} F_k\ (x_i) = y_i^{\wedge(t-1)} + f_t(x_i) \tag{4}$$

$$Obj(\theta)^{(t)} = \sum_{i=1}^{n} [g_i f_t\ (x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \gamma T + \frac{1}{2}\lambda \sum_{j-1}^{T} w_j^2 \tag{5}$$

From the above equation (6) describes the instance set of leaf J, from the tree structure q(x) contains the optimal leaf weight as w_j^2. The objective function is also described in equation (6) and (7).

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{6}$$

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \text{Y}T \tag{7}$$

The XGBoost training summary should be followed by a list of features and attributes. Features that contain the best splitting point for optimizing the training objectives.

3.1.4 Auto Encoder for UUD

The architecture in figure 4 describes the five phases that help claim the original news based on sentence-level features, pre-processing, data cleaning, and probabilistic deep learning. The CNN should contain the CNN layer, and the autoencoder should be encompassed by the MLP.

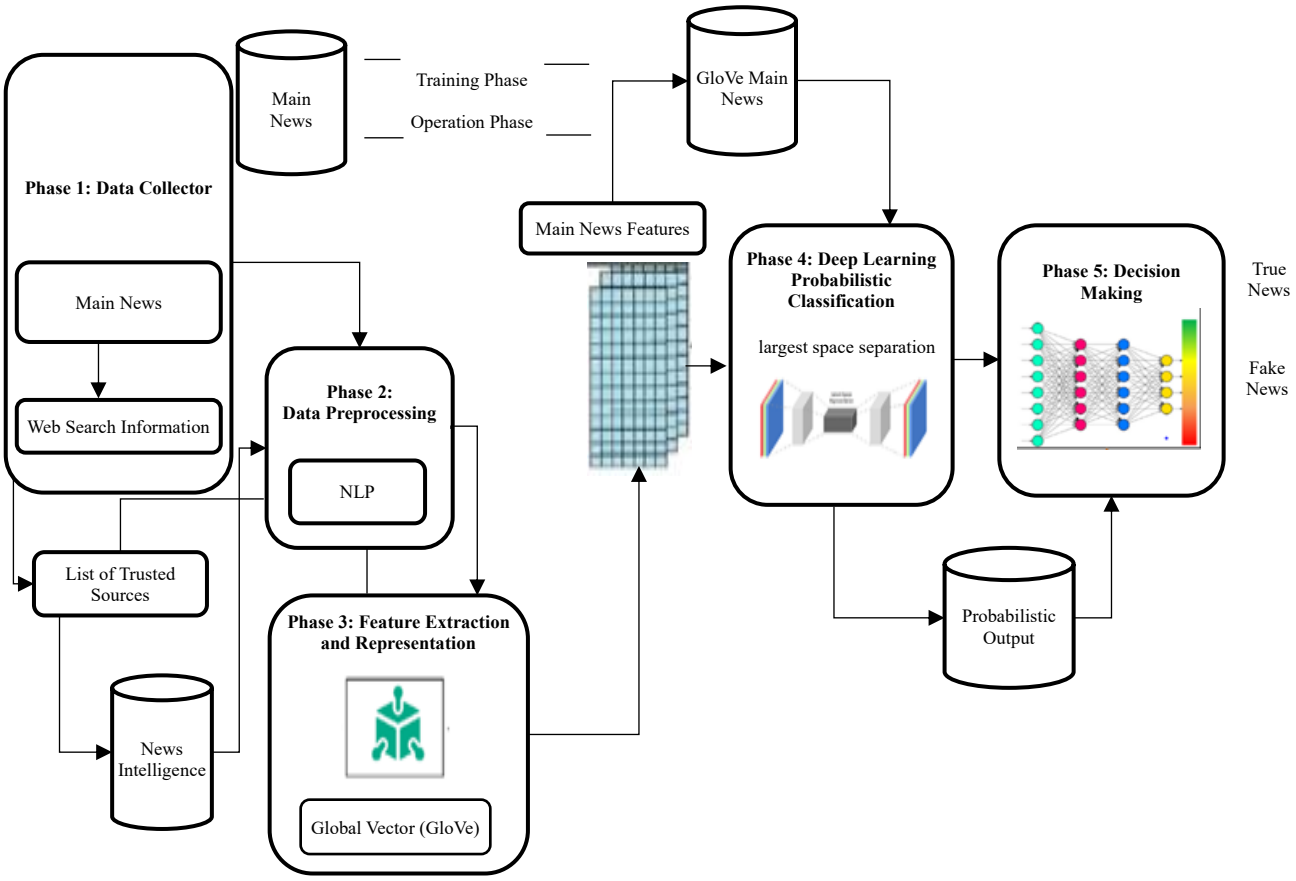

Figure 4: Auto encoder for UUD

Figure 4 shows the encoder and decoder algorithms for compressing the data. The AE approach is used for the detection method in the FAKE news case. The UFNDA autoencoder structure helps reduce the dimensionality of the feature space. The vector layer should be combined with the autoencoder to produce the input with the hidden states.

3.1.5 Decision Making

The MLP is an important regression and classification tool; it's used to assist in decision-making. The additional thing as a training feature should contain the MLP classified used the probabilistic output of $p(c)$. For the multi-class classification, the proposed model should have the MLP with four layers classified into two hidden layers and output layers. These layers vary in the neuron counts. Extract the input feature that contains the MLP classifier with the stacked MLP. The following is the formula for calculating the prediction of a given class, called p(c), for example, false news. The weight of neuron

i is represented by wi, the weights of the deep learners taught in the previous phase are represented by θ, and the matching output of the previous layer is determined by xi. If we want to know how well we did at predicting a specific class (like fake news), we may use the following formula to get p(c).

$$P(Class_{label} = c) = \sum_{i=1}^{n} w_{ci} * x_i + \theta \quad (8)$$

From the above equation (8) describes the logistic function followed by the predicted one.

$$logit\big(p(class_{label} = c)\big) = X_c = \log\left(\frac{P(class_{label} = c)}{1 - p(class_{label} = c)}\right) \quad (9)$$

From the above equation (9) describes about x_c logit is the logistic function of the predicted class, which contains the ranges of (−∞, +∞) with the ranges of [0,1]. The mentioned vector contains the probability that the input feature belongs to each class, and we find the maximum probability range to determine the class label.

$$Predicted\ class = index - of - \max(V) \quad (10)$$

From the above equation (10) describes the probability of input feature contains the class by finding the maximum probability, in this stage to determine the class label should be predicted.

3.1.6 SMOTE-Based Data Augmentation

To tackle the class-skew problem in our data, especially for hard-to-detect minority categories such as fake news, spam, and unauthorized users, we apply SMOTE (Synthetic Minority Over-sampling Technique) after the initial preprocessing pipeline. SMOTE works by generating synthetic samples for the underrepresented classes: it identifies nearest neighbours within each minority class, then interpolates between them to create new, plausible instances. This artificial balancing of the class distribution gives our classifiers a fairer training set, which typically translates into higher recall and a better F1-score for minority labels. In short, SMOTE levels the playing field so models can learn robust patterns across all classes, not just the majority ones.

Let the original dataset D be composed of Ninstances, where N_{min}instances belong to the minority class and N_{maj}instances belong to the majority class. The dataset imbalance ratio is given by:

$$\text{Imbalance Ratio} = \frac{N_{\text{maj}}}{N_{\text{min}}}$$

After applying SMOTE, the number of synthetic minority class samples generated is N_{synt}. The augmented minority class now has:

$$N_{\text{aug}} = N_{\text{min}} + N_{\text{synt}}$$

If we aim to achieve a 1:1 class distribution, the number of synthetic samples should satisfy:

$$N_{\text{synt}} = N_{\text{maj}} - N_{\text{min}}$$

Thus, the augmented dataset will have balanced classes, where:

$$N_{\text{total}} = N_{\text{maj}} + N_{\text{aug}} = 2N_{\text{maj}}$$

3.2 Proposed Algorithm

GBERT Algorithm for Fake News detection

Data: Labelled Text dataset

$Result: Binary\ fake\ news\ detection\ model$

$phase: 1\ Data\ Preprocessing$

$Dataset\ 1 \rightarrow collection\ (Text\ Dataset)$

$Dataset\ 2 \rightarrow Collection\ (Text\ Datset)$

Apply SMOTE on the training dataset to generate synthetic samples for the minority class.

$Phase: 2\ Model\ Generation$

$BERT$

$B_{embeddings} \rightarrow BERT(GBERT - dataset), used\ for\ bert\ base\ uncased\ pre\ trained\ model$

$B_{output} \rightarrow Pooled\ Output\ (B_{embeddings}) used\ for\ to\ extract\ BERT\ pooled\ output,$

$whcih\ is\ represents\ the\ entire\ input\ sequence.$

GPT

$G_{embeddings} \rightarrow G.decode\ (GBERT_{dataset}) - decode\ test\ using\ GPT$

$G_{extract} \rightarrow G\ last\_hidden\_state)$

$- Extract\ the\ final\ output\ representaion\ from\ the\ last\ hidden\ state.$

$Phase - 3: Dense\ net\ Processing\ Layer$

$Combine_{input} \rightarrow Merge(B_{Ouput}, G_{Output}) usedfor\ merge\ output\ from\ BERT\ and\ GPT$

$Processing\ layers.$

$Final_{Output} \rightarrow Dense.dropout(Combine_{input}), to\ combine\ input\ pass\ through$

$dense\ and\ dropout\ layers.$

$Binary - classification \rightarrow label\ (Final_{output}). Output\ is\ binary\ classiified\ as\ fake\ or\ real$

$XGBoost\ based\ Spam\ or\ Malicious\ Data\ Detetion\ Algorithm$

$Datset \rightarrow get\ labeled\ data\ for\ MDD(spam, legitimate\ messages)$

$Datset \rightarrow Preprocess\ data\ (remove\ noise, normalize\ features, encode)$

$X, Y \rightarrow extract\ features\ and\ labels\ from\ the\ datset$

$X = features\ (message\ content\ , user\ activity)$

$Y = labels\ \left(0 = legotomate, 1 = \frac{malicious}{spam}\right)$

$X - train, Y - train \rightarrow split\ datasets\ into\ training\ set$

$X - test, y - test \rightarrow split\ datsets\ into\ testing\ set$

$Initialize\ XGBoost\ model\ (XGBClassifier\ or\ XGBRegressor)$

$Train\ the\ XGBoost\ model\ using\ (X - train, y - train)$

$for\ i = 1,2,3, \ldots\ldots\ldots\ldots\ldots\ldots, n\ (test\ datset\ size)$

$prediction\ (i) \rightarrow XGBoost\ model.predict\ (X_{test(i)})$

$end\ for$

$calculate\ accuracy, precision, recall\ and\ F1\ Score\ to\ evaluate\ model$

$if\ prediction(i) == 1\ (malicious) then$

$x(i) identified\ as\ spam\ or\ malicious\ data$

$else$
$X(i) is\ identified\ as\ legitimate\ data$
$end\ if$
$re - sorted = sort(predictions, reverse = True)$
$\alpha \rightarrow get\ threshold\ from\ re - sorted\ where\ index = N - error - 1.$
$for\ i = 1\ to\ n$
$if\ prediction\ (i) > \alpha\ then$
$x(i) is\ spam\ or\ malicious$
$else$
$x(i) is\ legimate\ data$
$End\ if$
$end\ for$
$Unauthorized\ user\ detection\ Algorithm\ using\ Autoencoder$
$Input: The\ number\ of\ unauthorized\ user\ N - error$
$Output: Reconstruction\ error||x^1 - x||with\ unauthorized\ user\ detection$
$Dataset \rightarrow get\ user\ activity\ data\ from\ logs\ (login\ attempts, Ip\ address)$
$Dataset \rightarrow preprocess\ and\ normalize\ user\ activity\ data$
$x \rightarrow extract\ feature\ from\ user\ activity\ data\ using\ Dataset$
$X_{train} \rightarrow get\ normal\ user\ behavior\ dataset\ using\ X$
$X_{test} \rightarrow get\ test\ dataset(include\ both\ legitimate\ and\ unauthorozed)$
$Initialize\ autoencoder\ model\ with\ X_{train} data$
$Train\ autoencoder\ using\ X_{train} to\ learn\ normal\ user\ behaviour\ patterns$
$for\ i = 1,2,3 \ldots \ldots . . n(test\ dataset\ size)$
$Reconstruction\ error(i) = \left|\left|autoencoder\ \left(x(i)\right) - x(i)\right|\right|$
$end\ for$
$resorted = sort(reconstruction - error, reverse = True)$
$\alpha \rightarrow get\ threshold\ from\ resorted\ where\ index = N - error - 1$
$for\ i = 1\ to\ n$
$if\ reconstruction - error\ (i) > \propto\ then$
$x(i) is\ an\ unauthorized\ user\ (anomaly\ detected)$
$else$
$x(i) is\ the\ legitimate\ user$
$end\ if$
$end\ for$

4 Results and Discussion

4.1 Hardware and Software Configuration Details

Table 1: Hardware and software configuration details

Hardware Components	Specification
Processor (CPU)	Intel Core i7/i9 or AMD Ryzen 7/9
Graphics Processing Unit	NVIDIA GTX 1660 Super/RTX 2060/RTX 3060
Memory (RAM)	16 GB and 32 GB
Storage	500GB-1 TB SSD
Operating System	Windows 10/11, Linux or Mac OS
Power Requirements	Standard
Software Components	Specifications
Programming Language	Python 3.8+
ML/DL Frameworks	TensorFlow, Keras, PyTorch
Data Handling Libraries	Numpy, Pandas
Anomaly detection libraries	Tensor Flow Probability
Database Systems	MySQL, MongoDB
Environment Management	Anaconda, Virtualenv

Table 1 shows the hardware and software related to the proposed system. This also provides the important parts related to machine learning and deep learning models. It also carries 500 GB up to 1 TB of SSD storage to ensure data can be accessed fast and the loading time is reduced. It supports Windows 10/11, Linux, or even macOS, can be configured to be flexible between development platforms, but has conventional power requirements of normal workstation setups. Software development is based on machine learning and deep learning architectures such as TensorFlow, Keras, and PyTorch, and Python 3.8+ is the preferred choice of a fundamental language. The environment is managed using Anaconda or Virtualenv in order to achieve easy dependency management and installations and MySQL or MongoDB is utilized by the database layer in order to store structured and unstructured data. Broadly, such an arrangement ensures that the platform is well-built and refined to the extent that it is capable of supporting powerful AI-powered applications.

4.2 Performance Evaluation

4.2.1 Dataset Description

The proposed study combines multiple publicly available datasets including LIAR, ISOT, IFND, the Malicious Website URL dataset, and additional Kaggle-based sources such as the "Fake News Detection Dataset" (ISOT Kaggle version), the "Fake News Dataset" (commonly used for IFND-style multilingual misinformation), and the "Malicious and Benign URL Dataset" from Kaggle—to create a unified multi-domain detection framework. The LIAR dataset, sourced from PolitiFact, has approximately 12.8k brief political comments categorized in six honesty categories and detailed speaker metadata. The University of Victoria repository and Kaggle version of the ISOT dataset contain nearly 50,000 authentic and fake news items obtained of Reuters and unreliable online news portals on binary fake news classification. The Indian news websites, social media platforms, and public disinformation streams both with actual news and fake news are bilingual (English/Hindi), and both actual and fake news samples are shown in the IFND dataset, which is often uploaded in Kaggle under the names such as the Fake News Dataset (Indian news). The Malicious Website URL dataset, which is available on Kaggle as the Malicious and Benign URL Dataset, is a collection of threat data gathered by PhishTank, OpenPhish and

Malware Domains and harmless sources like Alexa top-ranked websites to offer lexical, domain, and host-level phishing URL details. The datasets are the basis of an Artificial Intelligence-driven Multi-Tiered Architecture that is supposed to detect fake news, spam/malicious material, and unlawful user behavior. The initial layer of the design is focused on data gathering and pre-processing, standardization of user behavior logs and URLs and textual content. Then it is followed by a feature engineering layer which derives TF-IDF vectors, word encodings, contextual encodings (including BERT and RoBERTa representations), URL lexical patterns, domain registration signals, and identity-based behavioural signals. The deep learning and machine learning algorithms used to classify spam URLs, disinformation, and abnormal user activity domains and used in the modeling tier are LSTM, CNN, BiLSTM, BERT, Random Forest, XGBoost, and one-class anomaly detectors. Finally, the decision and response layer offer real-time classification of fake news, spam filtering, block harmful URLs, and unauthorized user identification and a continuous learning layer makes model parameters to adapt to new information, feedback and new trends in threats. It is a multi-level, data-driven AI platform with an intelligent approach of detecting spam, web threats, fake news, and suspicious user activity under a unified system.

4.2.1.1 LIAR Dataset Sample Distribution

Table 2: LIAR dataset sample distribution

	training	Validation	Testing	Total
Pants fire (Ali et al., 2023)	839	116	92	1047
FALSE (Ali et al., 2023)	1994	263	249	2506
Barely true (Ali et al., 2023)	1654	236	212	2102
Half True (Ali et al., 2023)	2114	248	265	2627
Mostly True (Ali et al., 2023)	1962	251	241	2454
TRUE (Ali et al., 2023)	1676	169	207	2052
Total (Ali et al., 2023)	10239	1283	1266	12788

To interpret above table 2 represents the There are 12,788 cases in the LIAR dataset, which are divided into six truth categories: "Pants fire," "FALSE," "Barely true," "Half True," "Mostly True," and "TRUE." There are three groups of these instances: training (10,239 samples), validation (1,283 samples), and testing (1,266 samples). Out of all the categories, the greatest number of instances can be found in the "FALSE" category with 2,506 (1,994 in training, 263 in validation, and 249 in testing). Then there are 2,627 in the "Half True" category, 2,114 in the "Mostly True" category, and 251 in the "validation" and 241 in the testing categories. There are 2,052 instances in the "TRUE" category (1,676 in training, 169 in validation, and 207 in testing). There are 2,102 instances in the "Barely true" category (1,654 in training, 236 in validation, and 212 in testing). The "pants fire" category has the smallest representation, with 1,047 instances (839 in training, 116 in validation, and 92 in testing). This distribution guarantees that the dataset covers a wide range of truthfulness levels, with the training set being the most populated to allow for effective model learning, and the validation and testing sets providing appropriate representation of each category for model evaluation and performance assessment.

4.2.1.2 ISOT Dataset Sample Distribution

Table 3: ISOT dataset sample distribution

News Category	Training	Validation	Testing	Total
FAKE (Ali et al., 2023)	12850	2142	6425	21417
TRUE (Ali et al., 2023)	14089	2348	7044	23481
TOTAL (Ali et al., 2023)	26939	4490	13469	44898

Table 3 describes ISOT, which contains 44898 cases, divided into two divisions: FAKE and TRUE. It's also classified into training, testing, and validation. The training set should contain 26,939 samples, the validation set 4,490, and the testing set 13,469. Here, the TRUE set is defined as 23481, followed by the training as 14089, validation as 2142, and testing as 6425. If improving the training process means improving the model. If the validation and testing sets are not at a high level, it means that they have the best combination of FAKE and TRUE assessing models.

4.2.1.3 IFND Dataset Sample Distribution

Table 4: IFND dataset sample distribution

News Category	Training	Validation	Testing	Total
TRUE (Dhiman et al., 2024)	26460	3780	7560	37800

To interpret above table 4 represents the IFND dataset has 37,800 examples, and they all fall into the "TRUE" news category. These cases are separated into three categories: training, validation, and testing. The training set has 26,460 samples, the validation set has 3,780, and the testing set has 7,560. Most of the samples within the data set as "TRUE" are represented within the training sample and thus have a strong representation for the model to learn from, making them an ideal data sample. The validation and test samples, while smaller in quantity than the training sample still provide representative samples to assess the performance of the model and evaluate how the model will generalize to future unseen data samples.

4.2.1.4 Malicious Website URL Dataset Sample Distribution

Table 5: Malicious website URL dataset sample distribution

News Category	Training	Validation	Testing	Total
Malicious Website (Malik et al., 2024)	700,000	100,000	200,000	1,000,000

The Malicious Website URL dataset consists of one million entries where every entry falls under the category of "Malicious Website." Of these one million entries, the entries are categorized into three separate categories for the purposes of machine learning (ML) algorithm development: Training Set 700,000; Validation Set 100,000; Testing Set 200,000, as shown in table 5. As indicated in the table 5, most of the data allocated to the training set allows for greater learning capabilities for the ML model based on a larger number of examples. Even though there are fewer examples in the validation and test sets, the validation and test sets still contain sufficient information that will allow for a reasonable representation of what a potentially unsafe website may represent for the purposes of evaluating and testing performance of the machine learning model. The manner in which the datasets were created supports the ability of the machine learning model to be trained on and evaluated using multiple types of datasets, thereby ensuring its ability to generalize well to new malicious websites.

4.3 Performance Comparison of Various Models Using a Dataset (LIAR, ISOT, IFND and Malicious Website Dataset)

4.3.1 Performance Measure

To evaluate the proposed ideas for detecting fake news and calculate evaluation metrics, such as accuracy, recall, precision, and F1 score. These are used in the concepts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (11)$$

Equation (11) represents the detection accuracy among the model's performance metrics, i.e., true positives and true negatives. This does not define the costs of false negatives or false positives. Precision equation (12), which determines the number of false news, should be estimated by collecting a sample.

$$Precession = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (12)$$

The recall equation (13) is derived based on the number of correct classifications of false news divided by the amount of fake news in the dataset, as follows:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (13)$$

F1 Score can be calculated as,

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (14)$$

The F-measure, commonly known as the F1 score, is the overall performance when the dataset is unbalanced and false positives and false negatives are important, as equation (14) above explains. The F-measure ignores the true negative rate. Finding a balance between false positives and false negatives is crucial.

Table 6: Performance comparison of various models using a dataset (LIAR, ISOT, IFND and malicious website dataset)

Models	Dataset	Accuracy	Precision	Recall	F1score
LR (Nasir et al., 2021)	ISOT	55%	50%	52%	42%
RF (Nasir et al., 2021)	ISOT	96%	92%	92%	92%
MNB (Nasir et al., 2021)	ISOT	62%	60%	60%	60%
SGB (Nasir et al., 2021)	ISOT	55%	52%	52%	52%
KNNs (Nasir et al., 2021)	ISOT	61%	67%	61%	56%
DT (Nasir et al., 2021)	ISOT	98%	96%	96%	96%
AB (Nasir et al., 2021)	ISOT	94%	91%	91%	91%
CNN Only (Nasir et al., 2021)	ISOT	99%	99%	99%	99%
RNN Only (Nasir et al., 2021)	ISOT	98%	98%	98%	98%
Hybrid CNN-RNN (Nasir et al., 2021)	ISOT	99%	99%	99%	99%
RoBERTa (Ali et al., 2023)	LIAR	63%	62%	62%	62%
Conv-HAN (Ali et al., 2023)	LIAR	59%	59%	59%	59%
DEFNDM (Ali et al., 2023)	LIAR	51.05%	85.86%	45.47%	60.90%
EFNDM (Ali et al., 2023)	LIAR	51.05%	85.86%	45.47%	42.50%
ICNN-AEN-DM (Ali et al., 2023)	LIAR	89.59%	68.59%	69.09%	69.09%
XGBoost (Malik et al., 2024)	Malicious Webpages Dataset	86.60%	88.89%	82.64%	85.67%
BERT (Dhiman et al., 2024)	IFND	95.13%	95.21%	96.96%	96.08%
GPT-2 (Dhiman et al., 2024)	IFND	95%	94.11%	94.11%	94.08%
CNN+LSTM+LR (Dhiman et al., 2024)	IFND	94.19%	95.05%	95.54%	95.29%
GBERT (Dhiman et al., 2024)	IFND	95.30%	95.13%	97.35%	96.23%
MTA-SD	ISOT/LIAR/IFND/Malicious Webpages	99.90%	99.95%	99.96%	99.99%

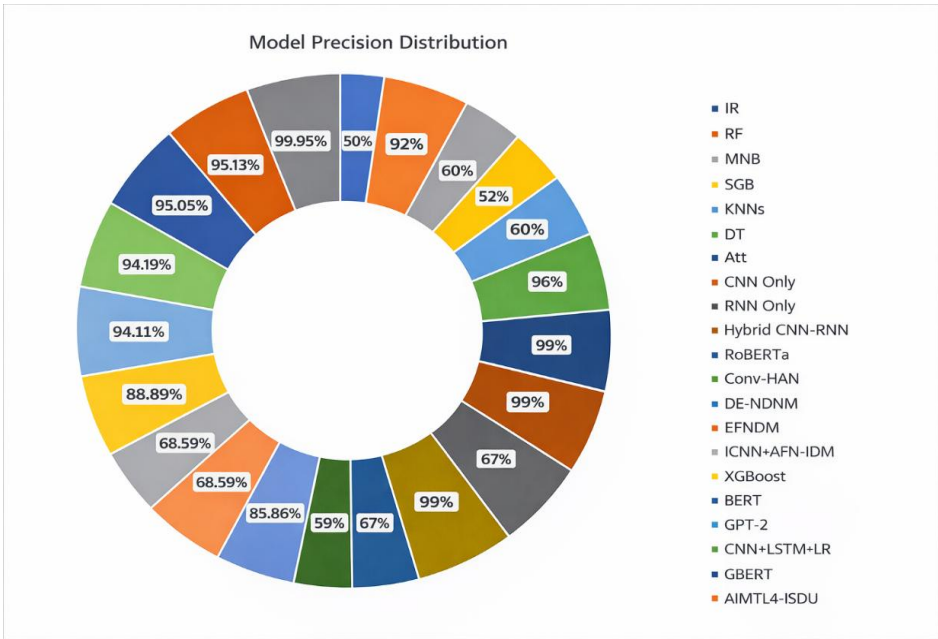

Figure 5: Precision vs. various models

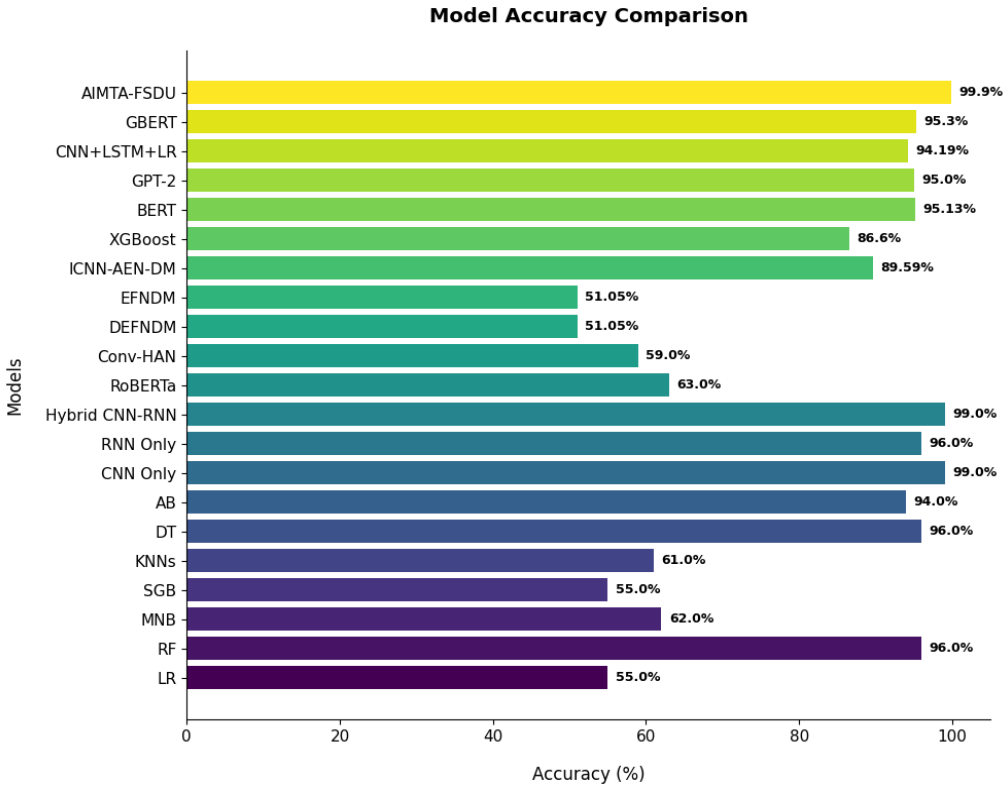

Figure 6: Accuracy vs various models

Figure 7: Recall vs various models

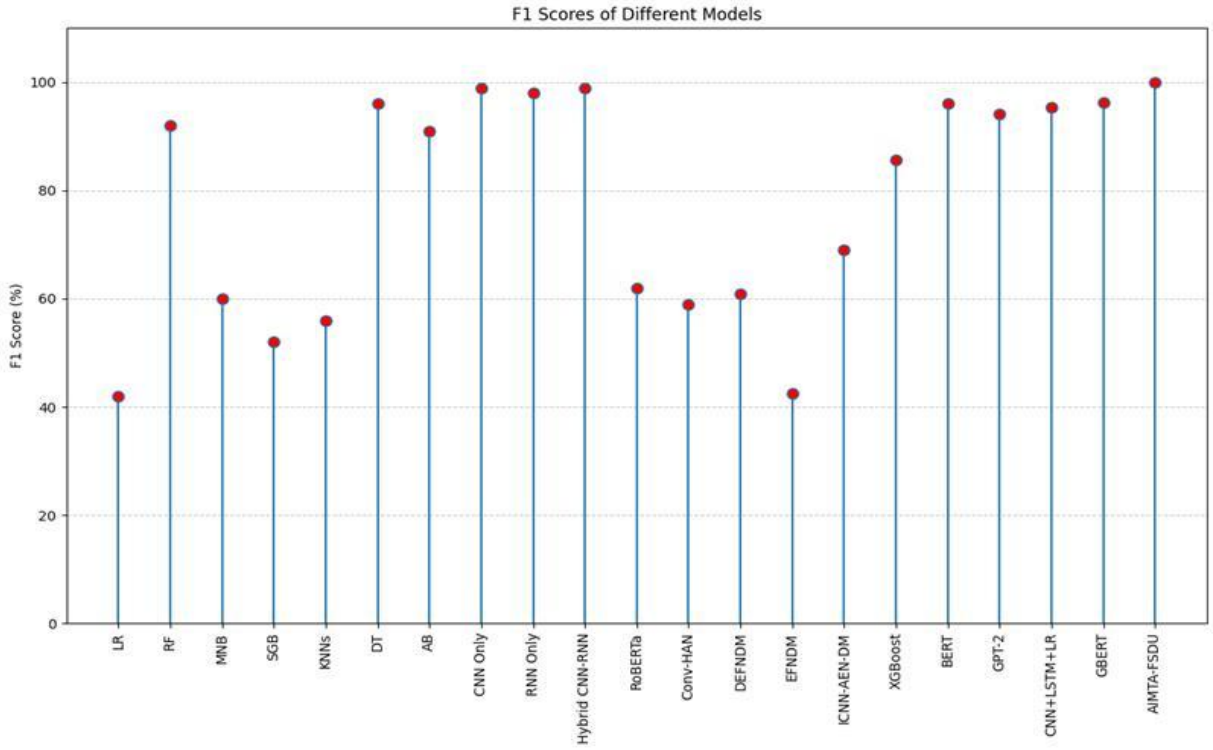

Figure 8: F1 score vs various models

The table 6 and figure (5), (6), (7) and (8) shows how well several machine learning models worked on distinct datasets: ISOT, LIAR, IFND, and Malicious Webpages. The metrics used to measure performance were Accuracy, Precision, Recall, and F1-score. A number of different traditional machine learning methods have been explored through Logistic Regression (LP), Random Forest (RF), Naive Bayes (MNB) and Decision Tree (DT) along with Deep Learning-based methods including CNN, RNN,

Hybrid CNN-RNN and Transformer based methods including BERT and GPT-2. The CNN-only and Hybrid CNN-RNN models obtained perfect data accuracy on the ISOT dataset at 99% while Decision Trees were the second-best performing algorithm (98%). ICNN-AEN-DM was the model class achieving the highest level of accuracy (89.59%) when tested on the LIAR dataset. The RoBERTa and Conv-HAN model classes did not perform quite as well based on the accuracy score that were obtained (63% for RoBERTa and 59% for Conv-HAN). Results from testing on the Malicious Webpages indicate that the XGBoost model (outclassing their rivals tested) again had very impressive accuracy results (86.60%). The MTA-SD model holds the best accuracy score out of tested models with very high accuracy (99.90%) when tested on IFND dataset. As for the Precision & Recall metrics of the tested algorithms they showed similar trends with MTA-SD leading the way (Precision 99.95% and Recall 99.96%) for all four datasets. Hybrid CNN-RNN models on ISOT and ICNN-AEN-DM models on LIAR performed well in both metrics, but models having DEFDNM and EFNDM applied to LIAR dataset had poor outcomes in both accuracy and the F1 score with results being around 51% accuracy. The F1 score is used to measure the balance of precision and recall. For example, MTA-SD has an exceptional F1 score of 99.99%. However, DEFDNM on LIAR received a low F1 score amounting to 42.50%. Generally, the results indicate that complex models such as MTA-SD, CNN architecture and transformer models like BERT and GPT-2 provide good F1 scores across all datasets for accuracy, precision, recall and F1 score. Conversely, simple models such as LR and SGB are hindered from performing well on certain datasets when it comes to obtaining a balanced measure of performance.

4.4 Discussion

The paper presents MTA-SD, a new Multi-Tiered Architecture developed within the AI paradigm to address issues related to fake news, spam, and unauthorized misuse of digital platforms. (ML), (DL), and (NLP) are used in a multi-layered solution to Fake News Detection (FND), Spam or Malicious Data Detection (MDD), and Unauthorized User Detection (UUD), all of which are tackled by MTA-SD. It operates on sophisticated algorithms such as BERT, GPT-2, and XGBoost to do efficient detections through textual data, contextual understanding, and user activity patterns. Compared to traditional models and deep learning systems, the model has a high success rate of 99.90% accuracy, 99.95% precision, 99.96% recall and an F1 score of 99.99% on different datasets. It is a powerful digital security tool as it is able to deal with multi-domain threats in real-time. The paper also gives recommendations on how this can be further improved in the future, including the use of multimodal learning to work with more complex data, cross-domain transfer learning to learn faster, real-time feedback loops, and enhancing interpretability to gain user trust. As well, it is possible to implement privacy-saving measures such as differential privacy so that the system can improve the security of data without affecting its performance. To sum up, the MTA-SD is an efficient and scalable tool to identify the digital threats, and there is a high possibility of future improvement to suit the arising challenges in the domain of digital security.

5 Conclusion

Due to the rapid increase in digital platform usage, fake news, spam, and illicit user activity detection have become a growing challenge. To assist in solving this issue, we have developed an MTA-SD (Multi-Tiered Architecture for Spam Data detection) system that combines advanced machine learning, deep learning and natural language processing methods into one integrated detection system. The MTA-SD will provide three detection functions for detection: FND - Fake News Detection; MDD - Spam/Malicious Data Detection; UUD - Unauthorized User Detection. The MTA-SD will utilize

cutting-edge technologies including BERT, GPT-2 and XGBoost, thus enabling it to deliver excellent results across a range of data types from text through to user logs/behaviour. In our initial trials across a range of datasets (ISOT, LIAR, IFND and Malicious Web Pages) the MTA-SD model produced outstanding results achieving 99.90% accuracy, 99.95% precision, 99.96% recall and 99.99% F1 Score. Furthermore, the multi-domain threat addressing capabilities of the MTA-SD model combined with its use of both hybrid artificial intelligence and advanced feature engineering techniques provide effective solutions for currently identified limitations of all other existing models including the limited scope and inability to generalize well across different datasets and issues of overfitting. There are multiple future enhancement opportunities for the MTA-SD model in terms of real-time threat detection and management. Its multimodal capabilities enhance the ability of the MTA-SD model to accommodate a wider variety of digital security risks (e.g., audio) and ultimately provide a more comprehensive set of capabilities. Second, by using cross-domain transfer learning, MTA-SD will have a reduced dependence upon large-scale labelled datasets; therefore, it will be able to transfer its knowledge and successfully apply it to different domains with less re-training than a model dependent on large-scale labelled datasets. Adaptable/feedback loops for real-time adaptation of MTA-SD can improve AIMTA's capacity to identify and mitigate emerging threats in real-time. Additionally, transparency and model interpretability will enhance the trust that stakeholders and users have in MTA-SD by enabling them to understand why AIMTA made a particular prediction or decision. Finally, in future incorporating privacy-preserving methods (e.g., differential privacy or federated learning) into MTA-SD would give MTA-SD the capacity to address digital security issues while ensuring user privacy. Ultimately, through use of these improvements, MTA-SD will be well prepared to meet the growing number of digital security challenges.

References

[1] Al Ghamdi, M. A., Bhatti, M. S., Saeed, A., Gillani, Z., & Almotiri, S. H. (2024). A fusion of BERT, machine learning and manual approach for fake news detection. *Multimedia Tools and Applications*, *83*(10), 30095-30112. https://doi.org/10.1007/s11042-023-16669-z

[2] Al-Haija, Q. A., & Droos, A. (2025). A comprehensive survey on deep learning-based intrusion detection systems in Internet of Things (IoT). *Expert Systems*, *42*(2), e13726. https://doi.org/10.1111/exsy.13726

[3] Ali, A. M., Ghaleb, F. A., Mohammed, M. S., Alsolami, F. J., & Khan, A. I. (2023). Web-informed-augmented fake news detection model using stacked layers of convolutional neural network and deep autoencoder. *Mathematics*, *11*(9), 1992. https://doi.org/10.3390/math11091992

[4] Al-Rawi, A., Groshek, J., & Zhang, L. (2019). What the fake? Assessing the extent of networked political spamming and bots in the propagation of fakenews on Twitter. *Online Information Review*, *43*(1), 53-71. https://doi.org/10.1108/OIR-02-2018-0065

[5] Babu, R., Kannappan, J., Krishna, B. V., & Vijay, K. (2023). An efficient spam detector model for accurate categorization of spam tweets using quantum chaotic optimization-based stacked recurrent network. *Nonlinear Dynamics*, *111*(19), 18523-18540. https://doi.org/10.1007/s11071-023-08697-z

[6] Balakrishnan, P., & Leema, A. A. (2025). Vulnerabilities and Defenses: A Monograph on Comprehensive Analysis of Security Attacks on Large Language Models. *Indian Journal of Information Sources and Services, 15*(2), 442–467. https://doi.org/10.51983/ijiss-2025.IJISS.15.2.54

[7] Chandrika, M. B., & Raju, A. N. (2025). Spammer Detection and Fake User Identification on Social Networks. *International Journal of Management Research and Reviews, 15*(2s), 377-386.

[8] Christy, C., Nirmala, A., Teena, A. M. O., & Amali, A. I. (2025). Machine learning based multi-stage intrusion detection system and feature selection ensemble security in cloud assisted vehicular ad hoc networks. *Scientific Reports*, *15*(1), 27058. https://doi.org/10.1038/s41598-025-96303-0

[9] Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, *10*(16). https://doi.org/10.1016/j.heliyon.2024.e35865

[10] Dubcek, A., & Voronina, E. (2021). The Role of Artificial Intelligence in Enhancing Human-Web Interaction. *International Academic Journal of Innovative Research, 8*(4), 1–5. https://doi.org/10.71086/IAJIR/V8I4/IAJIR0824

[11] Farokhian, M., Rafe, V., & Veisi, H. (2024). Fake news detection using dual BERT deep neural networks. *Multimedia Tools and Applications*, *83*(15), 43831-43848.

[12] Hossen, M. S., Sarker, M. T., Al Qwaid, M., Ramasamy, G., & Eng Eng, N. (2025). AI-driven framework for secure and efficient load management in multi-station EV charging networks. *World Electric Vehicle Journal*, *16*(7), 370. https://doi.org/10.3390/wevj16070370

[13] Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences*, *9*(19), 4062. https://doi.org/10.3390/app9194062

[14] Khalil, A., Hajjdiab, H., & Al-Qirim, N. (2017). Detecting fake followers in twitter: A machine learning approach. *International Journal of Machine Learning and Computing*, *7*(6), 198-202. https://doi.org/10.18178/ijmlc.2017.7.6.646

[15] Khule, M., Motwani, D., & Chauhan, D. (2025). A layered and integrative framework for Advance Persistent Threat detection and mitigation: combining AI, Zero-Trust, and Advanced Threat Intelligence. *Cluster Computing*, *28*(11), 740. https://doi.org/10.1007/s10586-025-05561-0

[16] Lim, K. S., Ooi, S. Y., Sayeed, M. S., Chew, Y. J., & Ahmad, N. M. (2025). Securing the Internet of Things: Systematic Insights into Architectures, Threats, and Defenses. *Electronics*, *14*(20), 3972. https://doi.org/10.3390/electronics14203972

[17] Lin, Y. (2024). Anomaly detection combining bidirectional gated recurrent unit and autoencoder in the context of e-commerce. *Engineering Research Express*, *6*(3), 035219. https://doi.org/10.1088/2631-8695/ad6819

[18] Malik, F., Suliman, M., Khan, M. Q., Rahman, N., Khan, K., & Khan, M. (2024). Optimizing malicious website detection with the XGBoost machine learning approach. *Journal of Computing & Biomedical Informatics*, *7*(02).

[19] Mokoena, G., & Nilsson, J. (2023). A Sophisticated Cybersecurity Intrusion Identification Model Using Deep Learning. *International Academic Journal of Science and Engineering, 10*(3), 17–21. https://doi.org/10.71086/IAJSE/V10I3/IAJSE1026

[20] Mounika, K., & Reddy, N. R. (2025). An integrated machine learning framework for spammer and fake user detection in online social networks. *Fringe Multi-Engineering Proceedings (FMEP)*, *1*(3), 12-25. https://doi.org/10.69996/fmep.2025015

[21] Naghib, A., Gharehchopogh, F. S., & Zamanifar, A. (2025). A comprehensive and systematic literature review on intrusion detection systems in the internet of medical things: current status, challenges, and opportunities. *Artificial Intelligence Review*, *58*(4), 114. https://doi.org/10.1007/s10462-024-11101-w

[22] Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International journal of information management data insights*, *1*(1), 100007. https://doi.org/10.1016/j.jjimei.2020.100007

[23] Park, S., & Choi, D. (2025). Exploring the potential of anomaly detection through reasoning with large language models. *Applied Sciences*, *15*(19), 10384. https://doi.org/10.3390/app151910384

[24] Qin, S., & Zhang, M. (2024). Boosting generalization of fine-tuning BERT for fake news detection. *Information Processing & Management*, *61*(4), 103745. https://doi.org/10.1016/j.ipm.2024.103745

[25] Rasul, H. M., & Jumaa, A. K. (2022). Real-time twitter data analysis: a survey. *UHD Journal of Science and Technology*, *6*(2), 147-155. https://doi.org/10.21928/uhdjst.v6n2y2022.pp147-155

[26] Saminathan, K., Mulka, S. T. R., Damodharan, S., Maheswar, R., & Lorincz, J. (2023). An artificial neural network autoencoder for insider cyber security threat detection. *Future Internet*, *15*(12), 373. https://doi.org/10.3390/fi15120373

[27] Sharma, S., Gaherwal, S., & Sharma, S. (2025). Real-Time Big Data Analytics in Social Media: Enhancing User Behavior Prediction. *AIJR Proceedings*, *7*(6), 155-170. https://doi.org/10.21467/proceedings.7.6.19

[28] Someswara Rao, C., Raminaidu, C., Raju, K. B., & Sujatha, B. (2024). Effective fake news classification based on lightweight RNN with NLP. Annals of Data Science, 11(6), 2141-2165. https://doi.org/10.1007/s40745-023-00506-z

[29] Soy, A., & Balkrishna, S. M. (2024). Automated detection of aquatic animals using deep learning techniques. *International Journal of Aquatic Research and Environmental Studies, 4*(S1), 1-6. https://doi.org/10.70102/IJARES/V4S1/1

[30] Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific reports*, *11*(1), 23705. https://doi.org/10.1038/s41598-021-03100-6

[31] Takiddin, A., Ismail, M., Zafar, U., & Serpedin, E. (2022). Deep autoencoder-based anomaly detection of electricity theft cyberattacks in smart grids. *IEEE Systems Journal*, *16*(3), 4106-4117. https://doi.org/10.1109/JSYST.2021.3136683

[32] Udayakumar, R., Joshi, A., Boomiga, S. S., & Sugumar, R. (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification. *Journal of Internet Services and Information Security, 13*(4), 138-157. https://doi.org/10.58346/JISIS.2023.I4.010

[33] Yi, J., Xu, Z., Huang, T., & Yu, P. (2025, February). Challenges and innovations in llm-powered fake news detection: A synthesis of approaches and future directions. In *Proceedings of the 2025 2nd international conference on generative artificial intelligence and information security* (pp. 87-93). https://doi.org/10.1145/3728725.3728739

[34] Zheng, Q., Zhao, P., Zhang, D., & Wang, H. (2021). MR-DCAE: Manifold regularization-based deep convolutional autoencoder for unauthorized broadcasting identification. *International Journal of Intelligent Systems*, *36*(12), 7204-7238. https://doi.org/10.1002/int.22586

Authors Biography

P. Kardeepa is currently working as an Assistant Professor in Department of Computer Science and Engineering, School of Computing at Kalasalingam Academy of Research and Education, KrishnanKoil. She is pursuing her Ph.D in Kalasalingam Academy of Research and Education, Deemed to be university, KrishnanKoil, Tamil Nadu. She has Qualified UGC NET and has 6 years of Teaching experience with more than 10 papers published in reputed Conferences. Her area of interest is Operating Systems, Context aware Computing and Machine Learning. She is a Professional member in ACM- Association for Computing Machinery.

N. Subbulakshmi is currently working as an Associate Professor in Department of Computer Science and Engineering, School of Computing at Kalasalingam Academy of Research and Education, Deemed to be university, KrishnanKoil. She has completed her post Doctorate at IIT Guwahati and her Research Area is Image and Signal Processing, Machine Learning. She has more than 15 years of Teaching experience and has published more than 35 papers in reputed Journals, National/International Conferences and Book Chapters and she has received Funding from various professional Schemes. Her area of interest is AIML, Computer Architecture, VLSI Design. She is presently supervising 6 Ph.D Scholars and She is a Professional member in ACM- Association for Computing Machinery and IEEE.

A.M. Gurusigaamani is currently working as an Assistant Professor in Department of Computer Science and Engineering, School of Computing at Kalasalingam Academy of Research and Education, KrishnanKoil. She is pursuing her Ph.D in Kalasalingam Academy of Research and Education, Deemed to be university, Krishnankoil, Tamil Nadu. She has more than 12 years of Teaching experience and has published more than 12 papers in reputed Conferences & journals. Her area of interest is Medical Image Processing, Deep Learning and Network security. She is a Professional member in ACM- Association for Computing Machinery & IEEE.