Deep learning-based multi-tier sensitivity analysis network for document sensitivity classification

Sadiya Ansari, Shameem Akther
Department of Faculty of Engineering and Technology, KBN University, Kalaburagi, India

Article Info

Article history:

Received Aug 21, 2024
Revised May 16, 2025
Accepted Jul 2, 2025

Keywords:

Document sensitivity classification
Multi-tier sensitivity analysis network
Encryption
Dual-scope graph convolution block
Deep learning

ABSTRACT

In the digital age, the exponential growth of data necessitates robust and efficient systems for document classification to maintain data security and compliance. Text classification plays a crucial role in identifying sensitive information by automatically categorizing documents based on their content. Using advanced machine learning and deep learning models, it analyzes text to detect keywords, patterns, and contextual cues that indicate the presence of sensitive data. This paper presents a novel framework, the multi-tier sensitivity analysis network (MTSAN), designed to accurately classify documents into public, private, and confidential categories. The proposed system integrates several advanced components, including the multi-tier sensitivity encoding network (MTSEN). MTSAN leverages a combination of convolutional networks and graph convolutional networks (GCNs) to capture both local and global contextual information. The dual-scope graph convolution block (DSGCB) is introduced to address both global dependencies and local dynamics, employing a novel fusion mechanism to merge global and local features effectively. Additionally, the cross-tier information fusion block (CTIFB) facilitates the seamless integration of multi-level features, further refining the classification process. The results demonstrate that the proposed MTSAN model outperforms traditional machine learning approaches and contemporary deep learning models such as bidirectional encoder representations from transformers (BERT), achieving superior accuracy and F1 scores in classifying sensitive information.

Corresponding Author:

Sadiya Ansari
Department of Faculty of Engineering and Technology, KBN University
Kalaburagi, India
Email: sadiyaansari_kbnu@rediffmail.com

1. INTRODUCTION

In today's digital era, the exponential growth of data has become both a boon and a challenge for organizations. As information proliferates, so does the need for effective data management and classification. The classification of documents into categories such as public, private, and confidential is essential for maintaining data security, privacy, and compliance with legal and regulatory frameworks. This necessity is particularly pronounced in sectors like finance, healthcare, and technology, where sensitive information must be meticulously protected to prevent unauthorized access and breaches. Ensuring the security and confidentiality of data to safeguard sensitive information from unauthorized access [1]. For data to be authentic, it needs to come from a trustworthy source and stay unchanged. Encryption and signature systems are essential for maintaining confidentiality and verifying authenticity. There are three types of data sensitivity public, private, and confidential. Public data refers to information that is openly accessible and poses minimal risk if disclosed. It includes content such as press releases, publicly available financial reports, and marketing materials. These documents are intended for wide distribution and do not require protection from unauthorized access. In contrast, private data includes information that is not meant for public disclosure but is not necessarily highly sensitive [2]. This category may encompass internal communications, personal opinions in customer reviews, and internal memos. While private data should be protected to maintain privacy, its unauthorized disclosure typically poses less risk than confidential information. Confidential data**,** on the other hand, includes highly sensitive information that, if exposed, could result in significant legal, financial, or reputational damage. This category covers a broad range of documents, such as medical records protected under laws like HIPAA (health insurance portability and accountability act), financial data, proprietary business information, and internal corporate communications [3], [4]. The protection of confidential data is paramount, as breaches can lead to severe consequences, including identity theft, financial loss, and loss of intellectual property.

Text classification, a subfield of natural language processing (NLP), plays a crucial role in automating the classification of documents into these categories. Text classification involves assigning predefined labels to text documents based on their content. Traditional methods relied heavily on manual review and rule-based systems, which are not scalable given the vast amounts of data generated daily. These methods often lack the flexibility and accuracy needed to handle complex and nuanced language in diverse documents. The advent of deep learning has revolutionized text classification, offering significant improvements in accuracy and efficiency. Deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures like BERT (bidirectional encoder representations from transformers), have demonstrated exceptional performance in understanding and processing natural language [5]. These models leverage large datasets and powerful computing resources to learn complex patterns and relationships within text, enabling them to classify documents with high precision. CNNs, typically used in image processing, have been successfully adapted for text classification by treating text as a sequence of words or characters, identifying local patterns, and combining them to form a comprehensive understanding. RNNs, and their variant long short-term memory networks (LSTMs), are well-suited for sequential data, capturing the context and dependencies within text. Transformer-based models, such as BERT, excel in handling long-range dependencies and context by using self-attention mechanisms, making them particularly effective for tasks involving nuanced language and contextual understanding [6]. The implementation of a robust document classification system not only enhances data security but also streamlines operations by ensuring that sensitive information is appropriately handled and accessed. It enables organizations to comply with data protection regulations, safeguard intellectual property, and maintain customer trust. As the landscape of data continues to evolve, the development and refinement of these classification systems remain a critical area of research and innovation [7].

In today's digital era, the rapid growth of data necessitates effective classification systems, particularly for sensitive information in sectors like finance, healthcare, and technology. Ensuring data security and privacy is crucial to prevent unauthorized access and breaches, while also meeting regulatory compliance requirements such as HIPAA and GDPR (general data protection regulation) [8], [9]. Traditional manual methods are insufficient to handle the sheer volume of data, prompting the need for automated, efficient, and accurate classification mechanisms. Recent advancements in deep learning and NLP, including CNNs, RNNs, and transformer-based models like BERT, offer promising solutions [10]. These technologies can enhance the accuracy and robustness of sensitive information detection, mitigating risks associated with misclassification and ensuring proper data handling. The motivation for this research is to leverage these technological advancements to develop a state-of-the-art system for categorizing documents into public, private, and confidential levels, thereby improving operational efficiency and compliance.

- Development of MTSAN framework: introduced the multi-tier sensitivity analysis network (MTSAN) for precise classification of documents into public, private, and confidential categories.
- Innovative multi-tier feature encoding: implemented the multi-tier sensitivity encoding network (MTSEN) to capture multi-tier document features, providing a nuanced understanding of content.
- Enhanced contextual understanding with DSGCB: developed the dual-scope graph convolution block (DSGCB) to effectively integrate global and local contextual information.
- Advanced feature fusion with CTIFB: utilized the cross-tier information fusion block (CTIFB) to seamlessly merge multi-level features, enhancing classification accuracy

2. RELATED WORK

The primary objective of this study is to develop a robust and efficient system for classifying documents into public, private, and confidential categories, particularly in the context of handling long and complex documents. This system aims to leverage advanced machine learning and deep learning techniques to accurately identify and differentiate sensitive information, thereby improving document management and ensuring data security. Pujar *et al.* [10] utilized BERT to segment lengthy texts, generating initial representations for each segment. The interactions between segments were subsequently modeled using either a recurrent layer or a transformer. Building on this, [11] introduced ERNIE-DOC, which includes a retrospective feed mechanism that allows for the integration of semantic information from the entire document. To better represent the structural features of long documents, [12] proposed the hierarchical graph convolutional networks (GCNs), constructing both section and word graphs to explicitly model macro and microstructural information. Additionally, sentence structures were incorporated into word graph modeling to enhance feature learning capabilities. To address the computational complexity inherent in the self-attention mechanism of transformers, [13] proposed various sparse attention mechanisms, aiming to restrict the range of token interactions and reduce computational overhead. Three main types of sparsity patterns have been identified: 1) fixed pattern, which includes window-based, global, and random attention techniques. For instance, [14] presented a hybrid approach that combines windowed local-context attention with task-driven global attention, effectively reducing computational complexity from quadratic to linear. This approach was further refined by [15], who incorporated random attention, maintaining linear complexity while improving model performance. 2) Learnable pattern, exemplified by [16], which dynamically determines the associated regions for each token, thereby enhancing the capture of semantic correlations. 3) Low-rank patterns, as explored by [17], where self-attention matrices are projected into lower-dimensional spaces to reduce complexity, leveraging the observation that these matrices often exhibit low-rank properties. Kitaev *et al.* [18] introduced the DeepDoc classifier, utilizing a deep CNN with the AlexNet architecture. This model was pretrained on the ImageNet dataset, significantly improving previous approaches by converting convolutional layers into flexible feature extractors.

Furthermore, Wang *et al.* [19] combined textual data from commercial OCR systems with raw image data. This data was then processed by a NLP model, which translated the text into the feature space. An extreme learning machine was employed to manipulate frozen convolutional layers trained with the AlexNet model, resulting in improved output without compromising accuracy. In the realm of hierarchical text classification, Martens and Provost [20] introduced the Seq2Label framework, which employs a random generative approach to learning label hierarchies by shuffling label sequences during training. Vermeire *et al.* [21] proposed the variational continuous label distribution learning (VCLDL) framework, which treats label distribution as a continuous density function in latent space. This method establishes a relationship between the feature and label spaces, uncovering information hidden in observable logical labels. Additionally, Lang *et al.* [22] introduced the prompt-based label-aware framework for multi-label text classification (PLAML). This framework enhances prompt-based learning with three key techniques: a token weighting algorithm considering label correlations, a template for augmenting training samples to make the process label-aware, and a dynamic threshold mechanism to refine the prediction condition of each label. Zhao *et al.* [23] further refined the multi-label classification approach by proposing PLAML, which specifically addresses the challenges associated with multi-label classification. These advancements collectively contribute to the development of more efficient and accurate document classification systems.

Despite significant advancements in document classification, existing methods, such as those leveraging BERT, ERNIE-DOC, and hierarchical GCNs, often face challenges in accurately categorizing sensitive information, especially in long and complex documents. The computational overhead associated with self-attention mechanisms and the limited ability to capture nuanced contextual relationships further exacerbate these challenges. Additionally, while frameworks like PLAML and VCLDL have improved multi-label classification, they may still fall short in scenarios requiring precise differentiation of document sensitivity levels [24], [25]. This research aims to address these gaps by developing a robust classification system that leverages advanced deep-learning techniques to enhance the accuracy and efficiency of identifying and categorizing sensitive information.

3. PROPOSED METHOD

This section of the proposed study focuses on the proposed model MTSAN that aims to classify sensitive data. This is categorized into various stages, namely private, public, and confidential stages having an increased accuracy. The proposed model has three major phases, the MTSEN, the proposed interactive network, and the integrated sensitivity feature fusion (ISFF). Figure 1 shows the MTSAN model.

The MTSEN consumes the input document and studies the multi-level expressions for sentences, words as well as sections. Further, the proposed “MTSAN” is implemented for data interaction between various levels of the convolutional network. Here, the convolutional network is designed particularly to handle lengthy documents, while the convolutional network for words as well as sentences is invested in grasping the microstructure of the document. The fine-tuned nodes that are redundant are omitted and the interaction of the multi-level data is improvised, section-aided segments for pooling as well as CTIFB are implemented prior as well as after every layer of the word as well as sentence convolutions. These are utilized to combine the three-level graphs into one integrated unit. We also introduce a global local graphical convolutional segment to dynamically grasp the local as well as the global features inside the sentence nodes which therefore increases the capabilities of feature representation. Lastly, the ISFF is implemented for integration of the prior multi-level features for the last of lengthy document categorization.

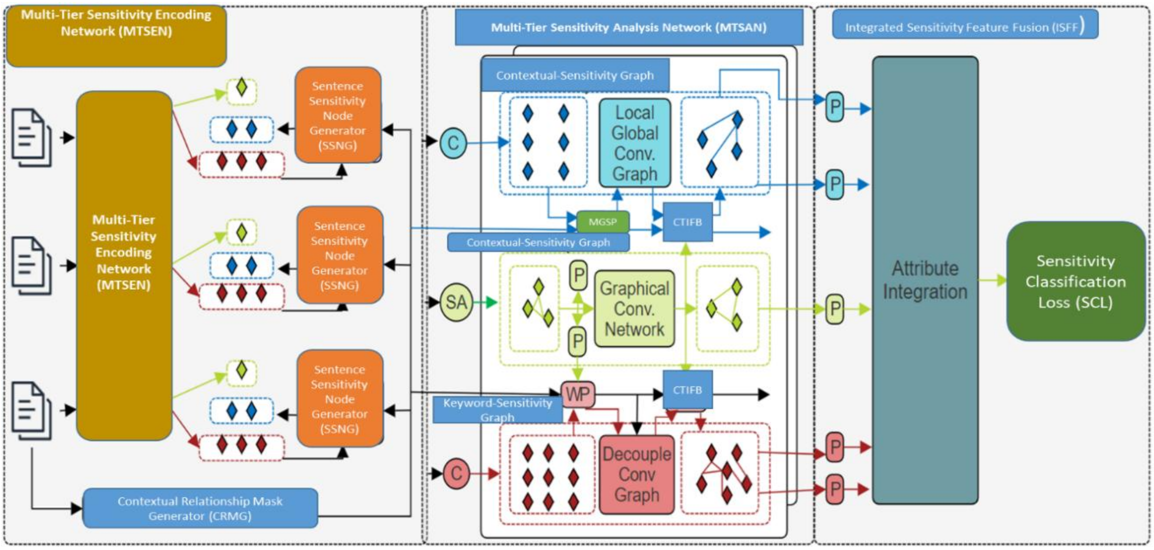
Multi-Tier Sensitivity Encoding Network (MTSEN)
Multi-Tier Sensitivity Encoding Network (MTSEN)
Sentence Sensitivity Node Generator (SSNG)
Sentence Sensitivity Node Generator (SSNG)
Sentence Sensitivity Node Generator (SSNG)
Contextual Relationship Mask Generator (CRMG)
Multi-Tier Sensitivity Analysis Network (MTSAN)
Contextual-Sensitivity Graph
C
Local Global Conv. Graph
MGSP
CTIFB
Contextual-Sensitivity Graph
SA
P
P
Graphical Conv. Network
WP
CTIFB
Keyword-Sensitivity Graph
C
Decouple Conv Graph
Integrated Sensitivity Feature Fusion (ISFF)
P
P
P
P
P
Attribute Integration
Sensitivity Classification Loss (SCL)

Figure 1. Multi-tier sensitivity analysis network

3.1. Multi-tier sensitivity encoding network

A document is normally made up of sections that could be further broken down into sentences and words. The document v is split into sections n for a fixed dimension to grasp the layered data representations omitting redundancy as is expressed as $\{r_1, r_2, \dots.., r_n\}$. Were, $r_k = \{y_{k0}, y_{k1}, \dots, y_{kt}\}$ is the $k-th$ section having $t+1$ tokens. To maintain a similar length t for every section, padding is utilized if needed. Every section is stored in the prior trained encoder $h(\cdot, \theta)$, the parameters here are denoted as $\theta.$ Here, in the last phase, the token x_k is considered a section attribute. Whereas, the other tokens are considered for word attributes. Further, resulting in section-level attributes for lengthy documents $X = [x_1, x_2, \dots, x_n]\ belongs\ to\ T^{n \times f}$ and for word level, it is given as $Z = [z_{11}, \dots.., z_{1t}, \dots, z_{n1}, \dots..., z_{nt}]\ belongs\ to\ T^{nt \times f}$. The attribute dimension is denoted as f.

The sentence expressions are attained by combining the word attributes for every sentence via pooling. Sentence masking $U_{masking}\ belongs\ to\ T^{nt}$ operation is used. Let us assume there are p sentences in the document, there are $i_k, k = 1,2, \dots, p$ words in the $k-th$ sentence. This is expressed as $U_{masking} = [1, \dots, 1,\ 2, \dots.., 2,\ 3, \dots.., 3, \dots.., p, \dots.., p]$. Here, the numbers that are identical showcase that the words at those locations are from the same sentence, the value of the numbers expresses the positional data of the sentences. Another projection layer is added to obtain the sentence attributes which are expressed as given in (1).

$$\underline{u}_k = \varsigma\ (Z[U_{masking} == k1_{nt}]), \text{ where } k = 1, \dots, p$$
$$u_k = Y_u \underline{u}_k + d_u,\ k = 1, \dots, p \qquad (1)$$

In this equation, we observe the max pooling is implemented column-wise. The trainable attributes are represented as $Y_u\ belongs\ to\ T^{f \times f}\ and\ d_u\ belongs\ to\ T^f$. In addition to sentence masking, there are two added transfer masking utilized in this network model that are denoted as $V_{u-x}\ belongs\ to\ T^{p \times n}$ and $V_{z-x}\ belongs\ to\ T^{nt \times n}$, which is used to establish the relationship between words, sentences as well as sections. The definition given for (2) is such that if a word or a sentence belongs to a particular section then the masking score is set to 1 otherwise it is set to 0.

$$[V_{u-x}]_{kl} = \{1,\ if\ the\ sentence\ k\ is\ from\ section\ l\ 0, else$$
$$[V_{z-x}]_{kl} = \{1,\ if\ the\ word\ k\ is\ from\ section\ l\ 0, else \qquad (2)$$

3.2. Multi-tier sensitivity analysis network

The proposed model consists of several graphical convolutional networks for words, and sentences as well as sections to grasp intra and inter-level correlations for lengthy documents. Considering the Section graph, the section scale attributes are given as $X\ belongs\ to\ T^{n\times f}$. Self-attention mechanism is implemented to build nodes for the completely linked graph I_x as given in (3).

$$I_x = (X, G, C_x), C_x = softmax\left(\left((XY^{Sx}) \times (XY^{Mx})\right)^V (f)^{\frac{-1}{2}}\right) \tag{3}$$

Here, the edge set is denoted as G, and the adjacent matrix is given as C_x for the graph I_x. Normalization which is row-wise is implemented and the weights are given as $Y^{Sx}\ belongs\ to\ T^{f\times f}, Y^{Mx}\ belongs\ to\ T^{f\times f}$. The dimension of the attribute is given as f. The output for the graphical convolutional network for the $q-th$ layer is as given below, where the diagonal node matrix is given as $\tilde{F}_{k,k} = \sum_l\ C_w^q(k,l). Y_w^q\ belongs\ toT^{f\times f}$ with μ as the activation function.

$$X^{q+1} = \mu\left(\frac{1}{\sqrt{\tilde{F}}}\frac{C_w^q}{\sqrt{\tilde{F}}}\ X^q\ Y_w^q\right) \tag{4}$$

Word, as well as contextual-sensitivity graphs, have nodes that begin with a huge quantity that has visible complexity while considering the computations involved. Therefore, we introduce section-aided pooling segments for sentences as well as section-aided pooling segments for words that allow pooling operations iteratively. This helps to omit nodes that are redundant as well as decreases the computational complexity and represents graphs for words and sentences at various levels. This improves the representation that is produced for the concluding attributes. We express the projection vector as $s = \varsigma(X^q)$, applied on the attributes of the $q-th$ layer of the macro-sensitivity graph, where $s\ belongs\ toT^f$. The index for the overhead nodes m_u^q is retrieved having scalar projection scores for sentences and similarly m_y^q for words. This is expressed as $a_u = \frac{U^q s}{\|s\|}$, $idz_u = ranking\ (a_u, m_u^q)$ and $a_y = \frac{Z^q s}{\|s\|}$, $idz_y = ranking\ (a_y, m_y^q)$. Here, the ranking of nodes is expressed as $ranking\ (a_u, m_u^q)$ and $ranking\ (a_y, m_y^q)$ that results in the largest m_u^q value in a_u and largest m_y^q value in a_y. The data propagation for section-aided pooling segments for sentences as well as section-aided pooling segments for words is formulated as given in (5).

$$\begin{aligned}&For\ Section\ aided\ pooling\ segments\ for\ Sentences,\\&\widetilde{U}^q = U^q(idz_u, :)\ \ V_{u-x}^{q+1} = V_{u-x}^q(idz_u, :)\\&\tilde{a}_u = sigmoid\big(a_u(idz_u)\big)\ \underline{U}^q = \widetilde{U}^q\ \otimes (\tilde{a}_u 1_f^V)\end{aligned} \tag{5}$$

$$\begin{aligned}&For\ Section\ aided\ pooling\ segments\ for\ Words,\\&\tilde{Z}^q = Z^q\big(idz_y, :\big)\ \ V_{z-x}^{q+1} = V_{z-x}^q(idz_y, :)\\&P_{z-u}^{q+1} = P_{z-u}^q\big(idz_y, idz_y\big)P_{z-x}^{q+1} = P_{z-x}^q\big(idz_y, idz_y\big)\\&\tilde{a}_y = sigmoid\left(a_y\big(idz_y\big)\right)\ \underline{Z}^q = \tilde{Z}^q\ \otimes (\tilde{a}_y 1_f^V)\end{aligned} \tag{6}$$

Considering the (4) and (5), the sub-matrices are $U^q(idz_u, :), Z^q\big(idz_y, :\big)$, $V_{u-x}^q(idz_u, :)$, $V_{z-x}^q(idz_y, :)$, $P_{z-u}^q\big(idz_y, idz_y\big)$ and $P_{z-x}^q\big(idz_y, idz_y\big)$ by choosing a row or column according to idz_u and idz_y. Element-based multiplication is represented as $\otimes$. Considering the contextual-sensitivity graph, we proposed a DSGCB segment to resolve the issue of unwanted data gathering by capturing semantic relations for various global areas. This resolution includes two convolutional graphs, one being global and the other one being local. The global graph is responsible for taking care of dependencies on long-term sentence nodes. Assume $\underline{U}^q$ is the pooled sentence attribute, the attention coefficients are evaluated with dot production for various nodes. Consider J as the number of nodes in the global graph. The attention matrix is denoted as $C_j = softmax\left(\frac{1}{\sqrt{f_m}}. \tilde{C}_j\right)$, here $f_m = \frac{f}{J}$. Therefore, we formulate as per given (7).

$$\tilde{C}_j = \underline{U}^q Y_{j,s}\big(\underline{U}^q Y_{j,m}\big)^V\ belongs\ to\ T^{m_u^q \times m_u^q} \tag{7}$$

Here, $Y_{j,m}$ and $Y_{j,s}\ belongs\ to\ T^{f\times f_m}$ are expressed as projection matrices. The output for the $j-th$ head is evaluated as $G_j = \mu(C_j \underline{U}^q Y_{j,x})$, where the matrix that is trainable is given as $Y_{j,x}\ belongs\ to\ T^{f\times f_m}$. The concluding integrated result is given as follows, with $Y_0 \; belongs \; to \; T^{m_u^q \times f}$ as the attribute matrix as given in (8).

$$G = Concatenate \left[\{G_j\}_{j=1}^{J}\right] Y_0 \; belongs \; to \; T^{m_u^q \times f} \tag{8}$$

While looking into the local convolutional graph segment, is developed to cater to the local dependencies considering the dynamic local data. An attention-masking window is also developed $C_n \; belongs \; to \, T^{m_u^q \times m_u^q}$, having a window dimension of ω, for shaping the relation between the nodes as well as their local locations. A threshold φ implemented on the attention head coefficients C_i for the global graph segment to choose the right semantic neighbor globally, which is expressed as given in (9).

$$[C_i\,]_{j,kl} = \{[C_i\,]_{j,kl}, \qquad [C_i\,]_{j,kl} \; greater \; than \; \varphi \; 0, \quad C_{j,kl} \; lesser \; than \; or \; equal \; to \; \varphi \tag{9}$$

Here, the $k - th$ and $l - th$ row as well as the column of the attention coefficient is expressed as $[C_i\,]_{j,kl}$. Lastly, the layer-based data propagation for the local convolutional graph segment, similar to the global convolutional graph segment is given as given in (10). Where the trainable parameters are given as ${Y_0}^*$ and ${Y_{j,n}}^*$. On the integration of the global as well as the local convolutional graph segments, we use an attribute fusion gate that is used in obtaining the concluding attribute representation for sentences. Therefore, we obtain the (11).

$$\begin{aligned} H_j &= \mu({C_{j,n}}^* \underline{U}^q {Y_{j,n}}^*), \quad j = 1,2, \ldots J \\ H &= Concatenate \left[\{H\}_{j=1}^{J}\right] {Y_0}^* \; belongs \; to \; T^{m_u^q \times f} \end{aligned} \tag{10}$$

$$\begin{aligned} I_u &= sigmoid(Y^{i1}[G;H] + d^{i1}), \quad j = 1,2, \ldots J \\ \vec{U}^q &= I_u \oplus G + (1 - I_u) \oplus H \end{aligned} \tag{11}$$

Here, the operation for concatenation element-wise is expressed as $\oplus$, $Y^{i1} \, belongs \; to \, T^{2f \times f}$ and $d^{i1} \; belongs \; to \, T^f$ are the learning variables. The words of the same section or sentence normally have more essential data. The layered structure of data is implemented for intra-section learning graphs to enhance the attribute interaction. C_y^q is used to represent the adjacent matrix that results from the operation of self-attention on the pooling word attribute $\underline{Z}^q$, is initially decoupled as an intra-section matrix that is also adjacent and expressed as C_{intra}^q inter section matrix C_{inter}^q for the length m of the section. The masking of sentences as well as sections are combined with C_{intra}^q to formulate the convolutional graph for intra-section as given in (12).

$$\begin{aligned} C_{prior}^q &= softmax\left(\alpha P_{z-u}^{q+1} + \delta P_{z-u}^{q+1}\right) \\ \hat{C}_{intra}^q &= softmax(softmax(C_{intra}^q) \oplus C_{prior}^q) \\ \underline{Z}_{intra}^q &= \mu(\hat{C}_{intra}^q \underline{Z}^q Y_{intra}^q) \end{aligned} \tag{12}$$

Where α and δ are used to denote hyperparameters. For inter-section convolutional graph as mentioned in (13). A Gate technique is applied between $\underline{Z}^q$ and $\vec{Z}^q$ that results in as given in (14).

$$\begin{aligned} \hat{C}_{inter}^q &= softmax(C_{inter}^q) \\ \vec{Z}^q &= \mu\left(\hat{C}_{inter}^q \underline{Z}^q{}_{intra} Y_{inter}^q\right) + \underline{Z}^q{}_{intra} \end{aligned} \tag{13}$$

$$\begin{aligned} I_y &= sigmoid\left(Y^{I^2} \underline{Z}^q + Y^{I^3} \vec{Z}^q\right) + d^{I^2} \\ Z^{q'} &= I_y \oplus \underline{Z}^q + (1 - I_y) \oplus \vec{Z}^q \end{aligned} \tag{14}$$

Here, the learning variables for the gate are expressed as Y^{I^2}, Y^{I^3}, and d^{I^2}. The attribute interaction between the macro-sensitivity graph, contextual-sensitivity graph as well as keyword-sensitivity graphs have to be integrated which is performed using a CTIFB. If a word or sentence node k is from a section l then as given in (15). Here, the learning variables of transfer fusion are denoted as $Y_{Transfer}^u$ and $Y_{Transfer}^y$. The detailed working of the proposed model is explained in the given Algorithm 1.

$$\begin{aligned} &U^{q}_{Transfer} = P^{q+1}_{u-x} X^{q+1} \\ &U^{q+1} = concatenate\left(U^{q}_{Transfer}, \vec{U}^{q}\right) Y^{u}_{Transfer} \\ &Z^{q}_{Transfer} = P^{q+1}_{z-x} X^{q+1} \\ &Z^{q+1} = concatenate(Z^{u}_{Transfer}, Z^{q'})\ Y^{y}_{Transfer} \end{aligned} \tag{15}$$

Algorithm 1. MTSAN

```
Input     Dataset $F = \{(v_k, a_k)\}_1^{Q_F}$, language model $h(\cdot,\theta)$ that is prior-trained and $n, o, t, \omega, Q$,
$m_u^q, m_y^q, q = 1, \dots, Q$
Output    Trained model $P$ that categorizes $H$ classes
Step 1    For every iteration $d = 1,2, \dots, V$ do
Step 2       Sample of a batch $Z_d$ from $F$
Step 3        Multi-tier sensitivity encoding network (MTSEN) part:
Step 4         Representation through words and sections
Step 5          Sentence representation using (1)
Step 6          SRMG using (2) and (3)
Step 7           Implement Multi-Tier Sensitivity Analysis Network (MTSAN)
Step 8           For every layer $q = 1,2, \dots., Q$ do
Step 9      Global convolutional network graph based on sections using (3) and (4)
Step 10     section aided pooling segments for sentences using (5) and (6)
Step 11   Global convolutional graph for contextual-sensitivity graph segment using (7)-(11)
Step 12   Global convolutional graph for keyword-sensitivity graph segment using (12)-(14)
Step 13     Interactions using transfer fusion with (15)
Step 14   End for
Step 15   Feature fusion part
Step 16   Feature fusion calculation using equation 16
Step 17   The Multi-Tier Sensitivity Analysis Network (MTSAN) is evaluated using (17)
Step 18   End For
Step 19   Return $P$
```

3.3. Integrated sensitivity feature fusion

A pooling operation based on the column is applied to gather more data at different layers which are expressed as $\varsigma(\cdot)$, the final output is the result of the evaluation as given in (16). The count of layers in the proposed model is indicated by Q. Then, another max pooling operation is performed to integrate the prior features of the layers which are $\breve{X}$, $\breve{U}$, and $\breve{Z}$ which is formulated as given below in (17).

$$\begin{aligned} &\breve{X} = \varsigma(X^{Q}) \\ &\breve{U} = [\varsigma(U^{1}), \dots., \varsigma(U^{Q}))] \\ &\breve{Z} = [\varsigma(Z^{1}), \dots., \varsigma(Z^{Q}))] \end{aligned} \tag{16}$$

$$w = \varsigma([\breve{X}, \breve{U}, \breve{Z}]) \tag{17}$$

4. PERFORMANCE EVALUATION

The paper introduces the MTSAN, a novel framework for classifying documents into public, private, and confidential categories. Leveraging advanced deep learning techniques, including convolutional networks and GCNs, MTSAN captures both local and global contextual information. It incorporates several key components, such as the MTSEN, DSGCB, and CTIFB, to enhance feature representation and classification accuracy. The proposed system was evaluated on datasets like the 20 newsgroups, enron email, and MIMIC-III clinical database, demonstrating superior performance over traditional models like support vector machine (SVM) and contemporary deep learning models like BERT. This study addresses existing gaps in sensitive information classification, particularly for complex and lengthy documents, by providing a more robust and efficient classification system.

4.1. Dataset details and comparison mechanism

4.1.1. 20 newsgroups dataset

This dataset consists of around 20,000 newsgroup documents, partitioned across 20 different newsgroups, covering a wide range of topics. The vast majority of the content in the 20 Newsgroups dataset can be classified as public. These are discussions from various newsgroups, covering a wide array of topics, intended for public viewing and sharing. There is minimal risk associated with this data being publicly accessible, as the content is meant to be open to all users of the newsgroups.

4.1.2. Enron email dataset

Many emails in this dataset can be considered private. As they involve internal corporate communications that are not meant for public disclosure. They do not necessarily contain highly sensitive information.

4.1.3. MIMIC-III clinical database

The clinical notes and medical data in the MIMIC-III database are highly confidential. This dataset includes sensitive patient information, medical histories, diagnoses, and treatment plans. The HIPAA in the United States, for instance, mandates strict confidentiality for such data to protect patient privacy.

4.1.4. Support vector machine

SVMs are a popular machine-learning technique used for classification tasks. They work by finding the optimal hyperplane that separates different classes in a high-dimensional space. In the context of document classification, SVMs are effective due to their ability to handle sparse data and high-dimensional feature spaces, typical in text-based applications.

4.1.5. Bidirectional encoder representations from transformers

BERT, a state-of-the-art deep learning model for NLP, leverages the transformer architecture. BERT is pretrained on vast amounts of text data and fine-tuned for specific tasks, making it exceptionally good at understanding context and semantics. Unlike traditional models, BERT reads text bi-directionally, considering the context from both preceding and following words in a sentence, thus capturing richer information

4.2. Results

Figure 2 compares the overall accuracy of three different models; SVM, BERT, and the MTSAN in classifying documents into public, private, and confidential categories. The SVM model achieves an accuracy of around 84%, indicating its basic capability to perform the classification task, though it struggles with more complex distinctions. The BERT model, leveraging deep learning techniques and transformer architecture, significantly improves accuracy to approximately 91%, demonstrating a better understanding of nuanced language and context. The MTSAN outperforms both SVM and BERT, achieving an accuracy of about 96%. This suggests that the PS model incorporates advanced features or optimizations, making it particularly effective at accurately categorizing documents across different sensitivity levels. The superior performance of the PS model highlights its potential as a reliable tool for sensitive data classification, aligning well to ensure precise identification and handling of sensitive information.

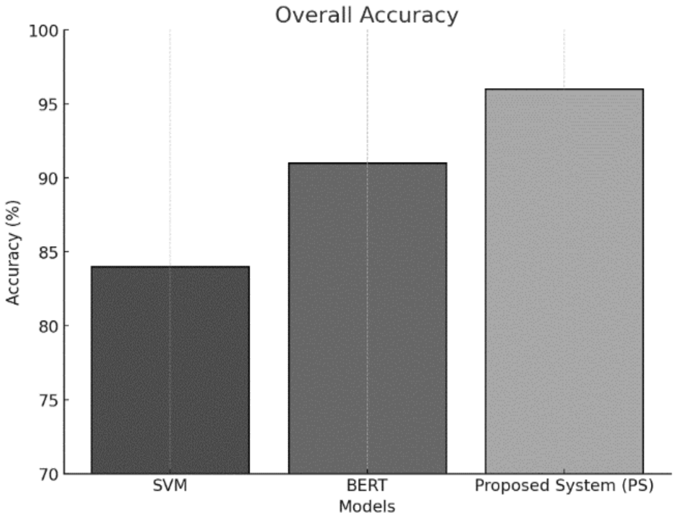

Figure 2. Compares the overall accuracy of three different models

Illustrates the F1-score for public data classification achieved by three models: SVM, BERT, and the MTSAN. The SVM model scores around 84.5%, indicating a relatively balanced precision and recall but with limitations in handling the complexity of public data content. The BERT model improves upon this, achieving approximately 91.5%, benefiting from its ability to understand context and semantics through deep learning. The MTSAN achieves the highest F1-score at around 96.5%, showcasing its superior capability in accurately classifying public documents. This high performance suggests that the PS model effectively balances precision and recall, minimizing both false positives and false negatives. The results highlight the PS model's advanced processing capabilities, making it the most reliable choice for categorizing public data within this set of models. Figure 3 compares the F1-score of public data for three different models.

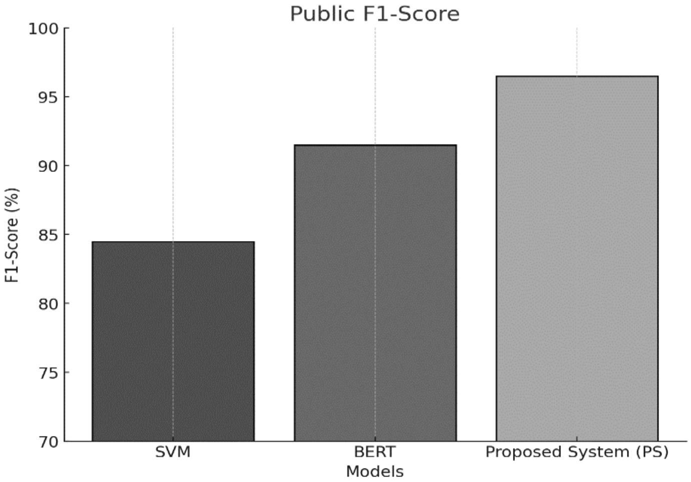

Figure 3. Compares the F1-score of public data for three different models

The bar chart displays the F1-score for classifying private data across three models: SVM, BERT, and the MTSAN. The SVM model shows an F1-score of approximately 82.5%, reflecting moderate performance in balancing precision and recall for private data, but with some shortcomings likely due to its less complex handling of nuances in the data. BERT, leveraging advanced NLP capabilities, achieves an improved F1-score of around 89.5%, indicating a more accurate classification of private documents, thanks to its ability to understand the context and intricate details. The MTSAN leads with the highest F1-score at approximately 94.5%, demonstrating exceptional precision and recall in identifying private information. This result indicates that the PS model is particularly effective in managing the complexities of private data, ensuring a high level of accuracy in classification, thus offering the most reliable performance among the evaluated models. Figure 4 compares the F1-score of private data for three different models.

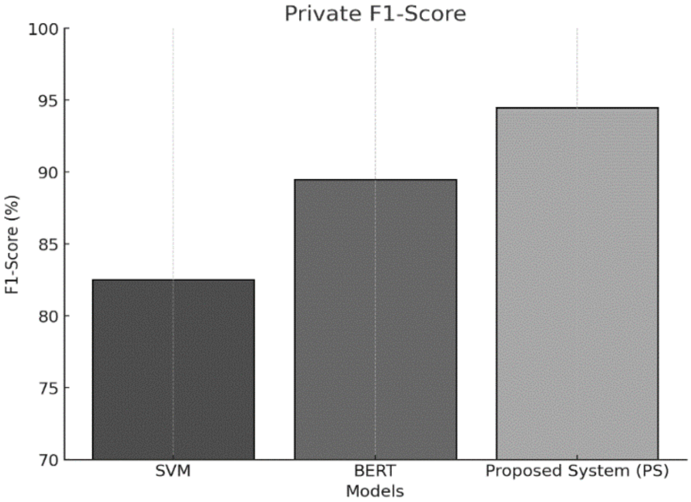

Figure 4. Compares the F1-score of private data for three different models

Figure 5 illustrates the F1-score for the classification of confidential data across three models: SVM, BERT, and the MTSAN. The SVM model has the lowest F1-score at approximately 79%, suggesting a struggle with accurately identifying and distinguishing confidential information, possibly due to its limitations in handling the subtleties of sensitive content. The BERT model improves on this with an F1-score of around 88%, benefiting from its deep learning architecture that better captures context and semantic nuances, thus enhancing its performance in classifying confidential documents. The MTSAN achieves the highest F1-score at about 93.5%, indicating superior precision and recall. This suggests that the PS model is highly effective at managing the complexities inherent in confidential data, ensuring a more accurate and reliable classification. The high performance of the PS model highlights its capability to maintain a balance between correctly identifying confidential information and minimizing false positives and negatives, making it the most effective model among those evaluated for this task.

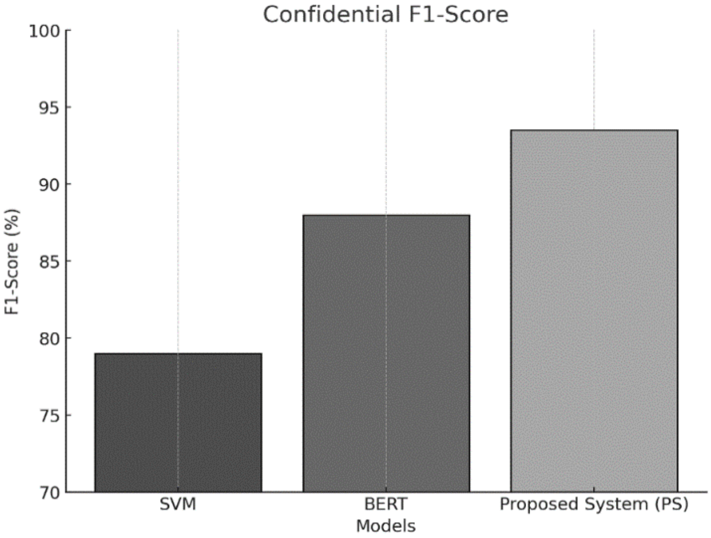

Figure 5. F1-score for the classification of confidential data

4.3. Comparative analysis

The comparative analysis demonstrates that the proposed MTSAN significantly outperforms both traditional machine learning models, such as SVM, and advanced deep learning models like BERT. MTSAN achieved a higher accuracy rate of approximately 96%, compared to 84% for SVM and 91% for BERT, showcasing its superior capability in correctly categorizing sensitive information into public, private, and confidential categories. The F1-scores also favored MTSAN, particularly in handling complex distinctions in confidential data, where it reached 93.5%, compared to BERT's 88%. Key improvements over BERT include the introduction of the MTSEN, which processes multi-level document features such as sections, sentences, and words, offering a more nuanced understanding of document structure. The DSGCB enhances semantic relationship detection by capturing both global dependencies and local dynamics, surpassing BERT's token-level attention mechanism. Additionally, the CTIFB in MTSAN effectively integrates multi-level features, a capability not present in BERT's architecture, thus leveraging both macro and micro-structural information for more accurate classification. Furthermore, MTSAN addresses the challenge of handling long and complex documents more efficiently than BERT, which can struggle with lengthy inputs due to its quadratic complexity. The section-aided pooling and feature fusion mechanisms in MTSAN facilitate the management of documents with mixed content types and varying sensitivity levels. Overall, these enhancements make MTSAN a more reliable and effective tool for sensitive information classification, providing a comprehensive and efficient solution for managing sensitive data.

5. CONCLUSION

The proposed MTSAN presents a robust and effective solution for classifying documents into public, private, and confidential categories. By leveraging advanced deep learning techniques, including convolutional networks and GCNs, MTSAN successfully captures both local and global contextual information, enhancing its ability to accurately differentiate between varying levels of sensitive information. The introduction of components such as the MTSEN, DSGCB, and CTIFB contributes to the model's superior performance. Empirical evaluations on diverse datasets, including the 20 newsgroups dataset, enron email dataset, and MIMIC-III clinical database, demonstrate that MTSAN outperforms traditional machine learning models like SVM and even state-of-the-art deep learning models like BERT. The proposed system consistently achieves higher accuracy and F1-scores, making it a reliable tool for sensitive data classification.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, we extend our heartfelt thanks to our guide for her unwavering guidance, invaluable insights, and encouragement throughout the research process.

FUNDING INFORMATION

No funding is raised for this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sadiya Ansari	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	
Shameem Akther								✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Author declares no conflict of interest.

DATA AVAILABILITY

Dataset is utilized in this research.

REFERENCES

[1] J. Fields, K. Chovanec, and P. Madiraju, “A survey of text classification with transformers: how wide? How large? How long? How accurate? How expensive? How safe?,” *IEEE Access*, vol. 12, pp. 6518–6531, 2024, doi: 10.1109/ACCESS.2024.3349952.

[2] A. Alabdulatif, I. Khalil, and X. Yi, “Towards secure big data analytic for cloud-enabled applications with fully homomorphic encryption,” *Journal of Parallel and Distributed Computing*, vol. 137, pp. 192–204, Mar. 2020, doi: 10.1016/j.jpdc.2019.10.008.

[3] J. Li, X. Kuang, S. Lin, X. Ma, and Y. Tang, “Privacy preservation for machine learning training and classification based on homomorphic encryption schemes,” *Information Sciences*, vol. 526, pp. 166–179, Jul. 2020, doi: 10.1016/j.ins.2020.03.041.

[4] B. Joshi, B. Joshi, A. Mishra, V. Arya, A. K. Gupta, and D. Peraković, “A comparative study of privacy-preserving homomorphic encryption techniques in cloud computing,” *International Journal of Cloud Applications and Computing*, vol. 12, no. 1, pp. 1–11, Sep. 2022, doi: 10.4018/IJCAC.309936.

[5] M. J. Zideh, P. Chatterjee, and A. K. Srivastava, “Physics-informed machine learning for data anomaly detection, classification, localization, and mitigation: a review, challenges, and path forward,” *IEEE Access*, vol. 12, pp. 4597–4617, 2024, doi: 10.1109/ACCESS.2023.3347989.

[6] S. Jiang, J. Hu, C. L. Magee, and J. Luo, “Deep learning for technical document classification,” *IEEE Transactions on Engineering Management*, vol. 71, pp. 1163–1179, 2024, doi: 10.1109/TEM.2022.3152216.

[7] K. Fiok *et al.*, “Text guide: improving the quality of long text classification by a text selection method based on feature importance,” *IEEE Access*, vol. 9, pp. 105439–105450, 2021, doi: 10.1109/ACCESS.2021.3099758.

[8] L. Meng, “The convolutional neural network text classification algorithm in the information management of smart tourism based on internet of things,” *IEEE Access*, vol. 12, pp. 3570–3580, 2024, doi: 10.1109/ACCESS.2024.3349386.

[9] J. Deepika, C. Rajan, and T. Senthil, “Security and privacy of cloud- and IoT-based medical image diagnosis using fuzzy convolutional neural network,” *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6615411.

[10] P. Pujar, A. Kumar, and V. Kumar, “Efficient plant leaf detection through machine learning approach based on corn leaf image classification,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, pp. 1139–1148, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp1139-1148.

[11] S. H. Sreedhara, V. Kumar, and S. Salma, “Efficient big data clustering using Adhoc Fuzzy C means and auto-encoder CNN,” in *Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems*, Springer, 2023, pp. 353–368.

[12] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, May 2019, vol. 1, pp. 4171–4186.

[13] S. Ding *et al.*, “ERNIE-DOC: a retrospective long-document modeling transformer,” in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, vol. 1, pp. 2914–2927, doi: 10.18653/v1/2021.acl-long.227.

[14] T. Liu, Y. Hu, B. Wang, Y. Sun, J. Gao, and B. Yin, “Hierarchical graph convolutional networks for structured long document classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 8071–8085, Oct. 2023, doi: 10.1109/TNNLS.2022.3185295.

[15] A. Vaswani *et al.*, “Attention is all you need,” in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[16] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: the long-document transformer,” *arxiv preprint: 2004.05150*, 2020, doi: 10.48550/arXiv.2004.05150.

[17] M. Zaheer *et al.*, “Big bird: transformers for longer sequences,” in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[18] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer: the efficient transformer,” in *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[19] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: self-attention with linear complexity,” *arxiv preprint: 2006.04768*, 2020, doi: 10.48550/arXiv.2006.04768.

[20] D. Martens and F. Provost, “Explaining data-driven document classifications,” *MIS Quarterly*, vol. 38, no. 1, pp. 73–99, Jan. 2014, doi: 10.25300/MISQ/2014/38.1.04.

[21] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, “Explainable image classification with evidence counterfactual,” *Pattern Analysis and Applications*, vol. 25, no. 2, pp. 315–335, May 2022, doi: 10.1007/s10044-021-01055-y.

[22] O. Lang *et al.*, “Explaining in style: training a GAN to explain a classifier in StyleSpace,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 673–682, doi: 10.1109/ICCV48922.2021.00073.

[23] X. Zhao, Y. An, N. Xu, and X. Geng, “Variational continuous label distribution learning for multi-label text classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 6, pp. 2716–2729, Jun. 2024, doi: 10.1109/TKDE.2023.3323401.

[24] X. Yan, H. Huang, Y. Jin, L. Chen, Z. Liang, and Z. Hao, “Neural architecture search via multi-hashing embedding and graph tensor networks for multilingual text classification,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 350–363, Feb. 2024, doi: 10.1109/TETCI.2023.3301774.

[25] T. Thaminkaew, P. Lertvittayakumjorn, and P. Vateekul, “Prompt-Based label-aware framework for few-shot multi-label text classification,” *IEEE Access*, vol. 12, pp. 28310–28322, 2024, doi: 10.1109/ACCESS.2024.3367994.

BIOGRAPHIES OF AUTHORS

Sadiya Ansari earned her Bachelors of Engineering BE degree in ISE from VTU, Belagavi in 2005. She has obtained her master's degree in M.Tech (CSE) from VTU Belgavi in 2012. And currently she is a research scholar at Khaja Bandanawaz University, doing her Ph.D. in computer science and engineering. She has 8+ years of teaching experience and has attended many workshops conducted by various universities. Her areas of interest are networking, machine learning, and artificial intelligence. She can be contacted at email: sadiyaansari_kbnu@rediffmail.com.

Dr. Shameem Akther is an Associate Professor in the Computer Science and Engineering Department at Faculty of Engineering and Technology, Khaja Bandanawaz University, Kalaburagi with an experience of 22 years in Teaching. She is qualified in Bachelor and Master Degrees in computer science and engineering, and Ph.D. in computer science and engineering in the area of image processing. Her areas of interest are data mining, computer vision, machine learning, artificial intelligence, internet of things, and image processing. She can be contacted at this email: shameemakther150@gmail.com.