

Deep Learning-based Detection of Depression Severity

Priyanka More¹, Rucha Patil², Sakshi Hedke³, Sanika patil⁴ and Juilee Talekar⁵

¹⁻⁵Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India
Email: priyanka.more@viit.ac.in, rucha.22210397@viit.ac.in, sakshi.22210064@viit.ac.in, sanika.22211056@viit.ac.in, juilee.22210007@viit.ac.in

Abstract— Identifying and assessing the severity of depression through social media text offers critical opportunities for enhancing mental health support systems. This paper conducts a comparative study of three deep learning approaches DeProBERTa and a hybrid CNN-BiLSTM model aimed at classifying depression severity from comments. A dataset comprising Reddit user posts was preprocessed through sentiment analysis using VADER, TF-IDF feature extraction, and text normalization techniques. Sentiment scores were mapped to severity labels, facilitating supervised model training. Among the models, the BiLSTM architecture demonstrated superior performance, achieving a validation accuracy of 84.10%, compared to DeProBERTa (70.36%). Explainable AI frameworks were utilized to interpret model decisions, highlighting influential linguistic patterns linked to depression levels. The study underscores the advantages of recurrent neural models in capturing complex emotional signals and stresses the role of interpretability in clinical natural language processing applications.

Index Terms— BERT, BiLSTM, DeProBERTa, Reddit Data, Sentiment Analysis.

I. INTRODUCTION

The increase in prevalence of mental health disorder with depression is a significant global concern that cuts across all age groups and social classes. Depression is characterized by sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, feelings of tiredness, and poor concentration, all of which may lead to severe impairment and limitation in the performance of one's daily routines and activities. According to the World Health Organization (2020), it is the leading cause of disability in the world and is a major contributor to the overall global burden of disease.

Traditional methods of depression diagnosis through clinical interviews and surveys like the Patient Health Questionnaire (PHQ-9) are usually reliable but may be time-consuming, expensive, and prone to the patient's desire to please or recall bias. In addition, a shortage of qualified mental health professionals around the world makes it difficult to diagnose and treat depression, especially in low and middle-income countries (Patel et al., 2018).

With the rapid growth and rise of social websites like Reddit, Facebook, and Twitter, a new era of making use of user generated text to infer the mental state has come into existence. This profound, real-time, and voluntary platform is suitable for people who wish to express their thoughts, emotions and discomforts more openly than in a clinical setting. The vast amount of textual data which is being generated on social media has made it a lucrative source for finding the first signs of depression.

In recent times, advancements in natural language processing (NLP) are observable. A case in point are deep learning models such as Bidirectional Encoder Representations from Transformers (BERT).

In the same vein, they come in handy in making it possible for a machine to comprehend the human language to a significant extent. Ideally, they delve deeper into the context of a piece of text, meaning that a machine can determine the context of a phrase. Ultimately, this is crucial in mental health-related tasks, which may involve the need to perceive the subtle expressions that are a common occurrence in most of the context switches. In this paper, we want to contribute to the classification of severity of depression by testing the performance of three state-of-the-art models: a fine-tuned DeProBERTa (optimized version of DeBERTa), the traditional BERT and a hybrid model CNN-BiLSTM to handle the task. In this regard, we carry out comprehensive preprocessing, sentiment analysis with VADER and feature extraction with TF-IDF, as well as a range of other methods to ensure high quality training. Moreover, we are well aware of the importance of transparency of AI in healthcare and, thus, employ a number of techniques to make sure the suggestions of the models are trustworthy and beneficial to the diagnostics, such as SHAP and LIME.

Efforts are being made in this work to address problems such as class imbalance, model overfitting, and interpretability in order to build robust, scalable, and interpretable models for assessing depression severity using social media data and help improve mental healthcare accessibility.

II. RELATED WORK

Depression is a major global health concern, and early detection through text-based analysis has gained significant interest among researchers. With the exponential growth of user generated data on social media platforms, natural language processing (NLP) techniques, especially using transformer based deep learning models like BERT and DeBERTa, have emerged as powerful tools for detecting depression severity. The arrival of BERT has revolutionized NLP by offering deep, bidirectional language modeling. In contrast to other traditional models, which read text from left to right, BERT's transformer architecture reads the entire text from one go and captures complex relationships in the process. This has a significant impact on mental health research, where subtle cues, unique metaphors, and especially sarcasm are common. Compared to ancient models, BERT-based classifiers perform much better in depression research [3].

According to the study, it was reported that instead of training the BERT model from scratch, fine-tuning the pre trained BERT model produces better results in depression detection. According to the study, it was reported that the F1-score, as well as the recall, improved by fewer epochs. The study also revealed that the general pre-trained BERT model was excellent, but fine-tuning for a specific job (like mental health-related issues) enhanced the performance better in picking up emotionally loaded language [4].

Fine-tuning of the BERT model is very helpful in emotion detection from text data DeBERTa is an improvement over BERT as it incorporates disentangled attention, meaning word content and positional encoding are separated. This architecture allows DeBERTa to model relationships between words regardless of their positions, which is essential in mental health texts as individuals may use creative or disordered syntax. DeBERTa has been shown to outperform BERT and RoBERTa for depression detection when properly fine-tuned [5].

Before deep learning algorithms gained traction and popularity, TF-IDF was considered the way to go when it comes to text representation in depression detection. The method is a way to show how significant a word is in a dataset based on how frequently it appears in a document. The importance is found by calculating how frequently a word appears in a document. This technique is helpful in traditional classifiers, such as SVM classifiers. However, deep learning algorithms have taken over, which is why we are using them. Despite this, TF-IDF is a good starting point, especially when resources are limited. [6]

BiLSTM networks are an improvement of the standard LSTMs in that they process the sequence forward and backward, hence capturing a lot of context, which is very critical in depression analysis. The importance is found by calculating how frequently a word appears in a document. since the meaning is most likely to depend on both prior and future words e.g. the statement "I am not happy" depends on the word "not" [7] and [11].

GloVe (Global Vectors for Word Representation) embeddings offer a compact word representation model based on statistical word-to-word co-occurrence calculations across a large text corpus. Combining GloVe with BiLSTM has been beneficial in the analysis of depression. This technique is helpful in traditional classifiers, such as SVM classifiers. However, deep learning algorithms have taken over, which is why we are using them. Despite this, TF-IDF is a good starting point, especially when resources are limited. Pretrained models provide useful semantic information to models that might take a large amount of labeled data to learn and improve performance and converge faster [8].

Mental health datasets are often imbalanced, with many more "Not Depressed" examples than "Severely Depressed" ones. Deep learning algorithms have taken over, which is why we are using them. Despite this, TF-IDF is a good starting point, especially when resources are limited. Techniques such as using weighted sampling,

oversampling the minority classes, and ensembling with a weighted loss function have been important to avoid biased models. Unbalanced datasets have tended to make models ignore minority classes, leading to poor recall rates in high-risk individuals' detection, which is a major failure in real-world applications [9].

While it is essential to have accuracy, it is also important for depression detection models to be interpretable. SHAP (SHapley Additive explanations) values tell us the contribution of each feature to the prediction, helping us understand why the model decided a comment was "Severe Depression". deep learning algorithms have taken over, which is why we are using them. Despite this, TF-IDF is a good starting point, especially when resources are limited. This is important for trustworthiness and the adoption of AI in mental health [10].

LIME is a distinct method for making predictions in machine learning. This method will help determine which words influenced a model to predict severe depression, making it possible for doctors and nurses to verify the AI model's output for each patient. The importance is found by calculating how frequently a word appears in a document. LIME stands for Local Interpretable Model Agnostic Explanations.

While transformers such as BERT and DeBERTa have become the industry standard, several comparative studies have shown that BiLSTM models – particularly when combined with CNN and enriched with GloVe embeddings – outperform transformers on smaller or noisier depression datasets. Recurrent models are excellent at capturing the sequence dynamics that are the key to depression language, especially when the data is sparse or fragmented [12].

III. METHODOLOGY

This section describes the comprehensive methodology used to identify depression severity in online posts using an advanced deep learning model. Two different neural architectures, DeProBERTa and BiLSTM, have been designed and finely tuned to classify reddit comments by four severity levels.

A. Data Preparation and Preprocessing

The data record consists of 39503 Reddit comments. Each comment comes with metadata that provides a context-related basis. The following preparation and processing steps have been performed to ensure high-quality input for the model, Honesty. Removing Null values: missing entries like Comments, missing text or related metadata, are removed to maintain dataset integrity. Text filtering: Data entries with fewer than five characters were removed to eliminate noise.

Lowercasing (Text Normalization): All text was standardized by converting to lowercase to ensure consistency in tokens.

Punctuation and Special Character Cleaning: Non-alphabetic characters were stripped to focus on textual content.

Tokenization: Comments were segmented into tokens using the NLTK tokenizer.

Stopword Removal: In order to preserve semantically meaningful terms, common stop words (such as "is", "the", and "at") were eliminated. Text data was transformed into numerical vectors using the TF-IDF (Term Frequency - Inverse Document Frequency) method: As features, the 1,000 most significant words were taken out. A TF-IDF matrix was generated, which transformed cleaned text and numerical representation

These preprocessing operations reduce noise and standardize input, ensuring that only linguistically and emotionally meaningful features are presented to the models.

B. Sentiment-Based Severity Labeling

Manual annotation of depression severity at scale is impractical. Hence, this study used VADER sentiment analysis to automate the labeling process.

Scoring: A compound sentiment score, with values ranging from -1, which indicates severely negative sentiment, to +1, which indicates extremely positive sentiment, was assigned to each examined comment. This approach allows objective and scalable annotation by aligning sentiment intensity with depressive indicators, suitable for downstream supervised learning.

TABLE I. SENTIMENT SCORE RANGES AND CORRESPONDING LABELS

Sentiment Score Range	Assigned Label
< -0.5	Severe(3)
-0.5 to 0	Moderate(2)
0 to 0.5	Mild(1)
> 0.5	Not Depressed(0)

C. Facial Expression Images Handling

In addition to 138,108 Reddit comments about mental health topics, the dataset also includes extra input data in the form of emojis and facial photos that reflect different emotional states. The model can learn from both visual signals and symbolic representations of emotions. The pre-processing procedures listed below are used to guarantee uniformity and quality among these input types.

Standardization of Input:

Emojis and Face Images: Emojis and facial photos are both converted into a single format so they may be added to the model simultaneously. Emojis are transformed into imagelike representations if they are in text format, and they are downsized to a uniform resolution (e.g., 224x224 pixels) if they include facial images.

Augmentation and Normalization:

Normalization: To ensure standardized input for training, the pixel values of both images—facial and emoji-based—are normalized to a range of [0, 1] or [-1, 1]. **Augmentation:** To provide variety and improve model resilience, data augmentation techniques (such as rotations, flipping, and color variations) are used for facial images. Emojis can also be altered (rotated, scaled) to mimic variations in input data found in the actual world.

Feature Extraction and Fusion: To create feature representations, each form of input is transmitted through the appropriate feature extraction layer [20][21][22][23] i.e., CNNs for input based on images. After that, these characteristics are merged into a single vector, which is sent into the model's last layers for classifying the degree of emotion or depression.

The above method is implemented if facial expression image or any emoji is present in the input data.

D. Feature Representation

For different architectures, the representation of textual features varied:

DeProBERTa: Input comments were tokenized using the DeBERTa tokenizer, ensuring each sequence was padded or truncated to 256 tokens, including attention masks for input encoding.

BiLSTM: Tokenized text was converted to vectors using pretrained GloVe embeddings (100-dimensional), capturing semantic and emotional nuance. Sequences were padded for uniformity.

Explanation: While transformers learn contextual embeddings during training, BiLSTM benefits from external semantic vectors like GloVe to encode word relationships effectively.

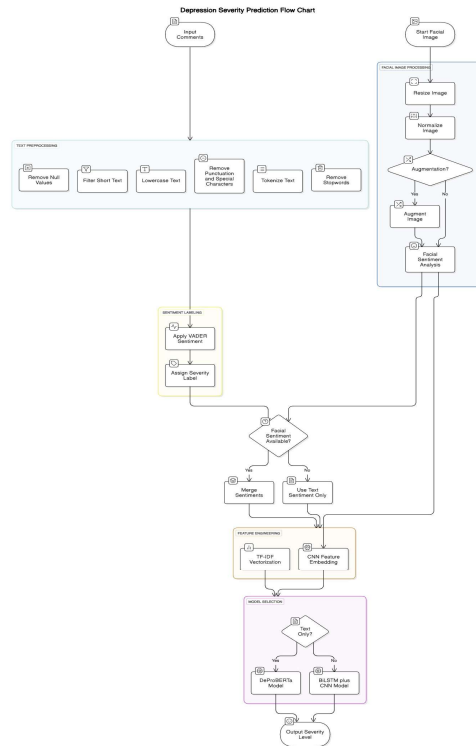


Fig. 1. Workflow Diagram

Fig.1 present here represents the workflow of entire system from taking input of text or emoji to assigning severity score to identify stage of depression with detailed stepwise breakdown. Each block represents its detailed working and flow.

E. Deep Learning Architectures

DeProBERTa-Based Model: The DeProBERTa model, a variant of RoBERTa with disentangled attention, processes position and content representations separately for enhanced context comprehension. DeProBERTa leverages rich contextual modeling, essential for discerning subtle emotional cues that indicate varying depression severity levels. Architecture Highlights:

TABLE II. RAINING OF DEPROBERTA

Parameter	Value
Epochs	5
Learning Rate	2e-5
Batch Size	16
Optimizer	AdamW
Sequence Length	256 tokens
Loss Function	CrossEntropyLoss
Precision Mode	Mixed (torch.cuda.amp)

Pretrained Backbone: DeProBERTa-v3-base

Transformer Layers: 12 layers with disentangled attention

Freezing Strategy: First 6 layers frozen; last 6 layers fine-tuned

Classification Head: Fully connected layer followed by Softmax

Training Configuration:

BiLSTM-CNN Hybrid Model: This model integrates CNN layers for local pattern detection and Bidirectional LSTMs for learning long-range dependencies in text sequences.

Architecture Highlights:

Embedding Layer: 100D GloVe embeddings with finetuning enabled

CNN Layer: 128 filters, kernel size = 3 (captures trigram sentiment cues)

BiLSTM Layer: 2 layers, bidirectional, hidden size =128, Dropout: 0.5

Layer Normalization: Applied before the final classification head

Final Layer: Fully connected + Softmax for 4-class output

Training Configuration: The model was trained with carefully chosen hyperparameters to ensure optimal performance. A total of 15 epochs was used to allow sufficient learning without overfitting. A learning rate of 2e-4 was established, providing a balance between convergence speed and stability during optimization. A Batch size of 32 was selected to effectively utilize the memory of GPU while maintaining stable gradient estimates.

The AdamW optimizer was employed, which is a variant of Adam that decouples weight decay from the gradient update, leading to better generalization. The loss function used was Weighted CrossEntropyLoss, which gives several classes varying weights, aiding in addressing class imbalance.

The hybrid design allows the CNN to extract emotionally charged n-gram features (e.g., "so tired"), while BiLSTM captures the contextual progression, improving severity classification.

F. Model Training and Evaluation Strategy

Both the DeProBERTa and BiLSTM models were trained and evaluated using a stratified data splitting approach, where 10% of the dataset was set out for validation and 90% was set up for training. Stratified sampling ensures that the distribution of depression severity classes remains proportionally consistent across both sets. This technique is critical to prevent skewed learning, particularly when dealing with class imbalance issues inherent in real-world social media data [24].

To thoroughly assess model performance, multiple Evaluation Metrics were used:

Accuracy: Measures the percentage of total correct predictions relative to all predictions made. It offers a broad summary of the model's performance. but it may not fully reflect the behavior of minority classes in imbalanced datasets.

Precision: Quantifies the proportion of positive samples that were accurately predicted among all positive forecasts. High precision means that the model is typically right when it predicts a particular class (e.g., severe depression).

Recall: Also called sensitivity, recall calculates the percentage of real-world present-positive situations that the model accurately detected. In jobs pertaining to healthcare, high recall is essential, where missing true positive cases (e.g., severely depressed individuals) can have serious implications.

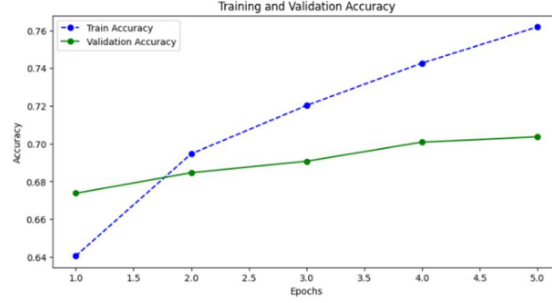


Fig. 2. DeProBERTa Accuracy plot

IV. RESULT AND DISCUSSION

In this section, we explored the experimental analysis of training and validating the DeProBERTa and BiLSTM models for classifying depression severity. We provide a thorough evaluation using various metrics to give a clear comparison of how each model performed.

A. DeProBERTa Performance

The DeProBERTa model, which was fine-tuned using Reddit comments, showed a steady improvement during its training. Over five training sessions, the accuracy steadily climbed, reaching 76.18% by the end. Meanwhile, the validation accuracy hit a high of 70.36 %. The model maintained a good balance between precision and recall across different classes, although there were some slight signs of overfitting after the third training session, as indicated by a slight increase in validation loss. This behavior suggests that while DeProBERTa effectively captures complex semantic patterns related to emotional states, it remains sensitive to limited fine-tuning data, requiring careful regularization.

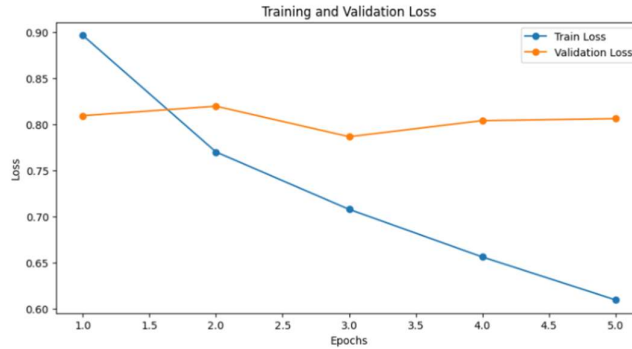


Fig. 3. DeProBERTa Loss Plot

B. BiLSTM-CNN Model Performance

The CNNBiLSTM is a hybrid model that outperformed the DeProBERTa model with all evaluation metrics. The training accuracy steadily increased during the 15 epochs, reaching 94.94%, while the validation accuracy stabilized at 84.10%. The BiLSTM model demonstrated strong performance in text classification, achieving high accuracy and balanced classification metrics across different severity levels. Training and Validation Accuracy: The accuracy curve indicates a steady improvement over the epochs. The final training accuracy reached 94.94%, while validation accuracy stabilized at 84.10%, reflecting robust learning. The small gap between Training and validation accuracy suggests a minute overfitting trend, but the model maintains strong generalization capabilities. The loss curves exhibit a consistent decrease in training loss, confirming effective optimization



Fig. 4. BiLSTM Accuracy Plot

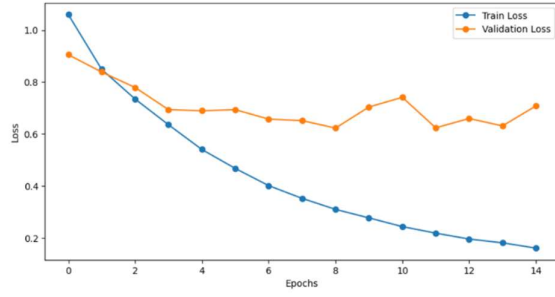


Fig. 5. BiLSTM Loss Plot

Validation loss follows a downward trend initially but fluctuates in later epochs, indicating a potential early onset of overfitting. This suggests that techniques such as dropout regularization or early stopping could further enhance stability.

The model demonstrates high classification accuracy, particularly for Class 1 (5348 correct predictions), Class 0 (4834), Class 3 (4892), and Class 2 (4457). These results suggest that the model effectively distinguishes between these categories. However, notable misclassifications are observed, such as 1,363 instances of class 3 being wrongly classified as class 0 and 1,018 samples from class 2 misclassified as class 1. Such errors indicate feature overlap between these classes, which may be addressed through enhanced feature selection and model optimization.

Classification Performance: The model achieves a balanced performance across all severity classes, with high precision, recall, and F1-scores. The results indicate that the model effectively differentiates between categories, particularly excelling in classes 1 and 3, which have the highest F1-scores.

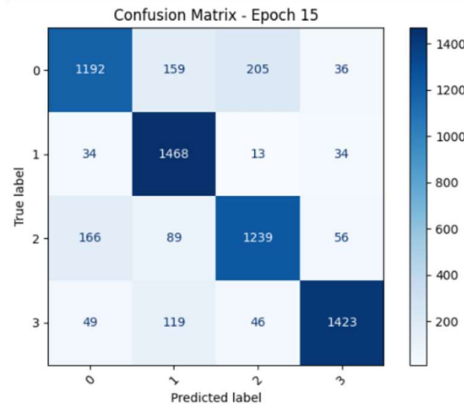


Fig. 6. BiLSTM Confusion Matrix

V. CONCLUSION

This study explored the application of deep learning techniques, specifically DeProBERTa and a CNN-BiLSTM hybrid model, to classify depression severity from social media posts. Through careful preprocessing, sentiment-based labeling, and robust feature extraction, both models demonstrated the ability to capture complex emotional and semantic signals from textual data. Among the two, the CNN-BiLSTM model performed notably better, achieving a validation accuracy of 84.10% compared to DeProBERTa's 70.36%. Its strength lies in combining local pattern detection through CNNs with the contextual memory of BiLSTMs, effectively identifying subtle emotional cues in text. Additionally, the use of explainability methods improved model transparency, which is crucial for mental health applications.

However, challenges such as misclassification between adjacent classes and minor overfitting were observed. Future work should aim at refining feature selection, enhancing regularization, and potentially incorporating multimodal data sources. Overall, this research highlights the potential of deep learning combined with explainable AI in developing reliable, scalable tools for mental health assessment.

REFERENCES

- [1] Bokolo, B. G., & Liu, Q. (2024). Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. *Electronics*, 13(20), Article 3980.
- [2] Qasim, A., Mehak, G., Hussain, N., Gelbukh, A., & Sidorov, G. (2025). Detection of depression severity in social media text using transformerbased models. *Information*, 16(2), Article 114.
- [3] Juarto, B. (2024). Sentiment analysis for mental health using boosting, bagging, and DeBERTa. In 2024 IEEE International Conference on Consumer Electronics (ICCE). IEEE.
- [4] Wang, N., Chiong, R., Kamil, R., & Zhang, W. (2024). Depression detection using speech audio and text: A comprehensive review. Authorea.
- [5] Kumar, R., Maharaj, K., & Saxena, A. (2024). Mental disorder classification via temporal representation of text. *arXiv preprint arXiv:2406.15470*.
- [6] Garc'ia-D'iaz, J. A., Pan, R., & Salmeron-R'ios, A. (2024). Fine-grain emotion analysis using linguistic features and transformers. *PeerJ Computer Science*, 10, e1992.
- [7] Nabilah'Izzaturrahmah, A., Sutrisno, S., Sutoyo, & Maimunah, S. (2024). Tweet-based depression detection using BERT optimized by grey wolf optimization. In *Proceedings of the 2024 IEEE International Conference*. IEEE.
- [8] Akintoye, O., Wei, N., & Liu, Q. (2024). Suicide detection in tweets using LSTM and transformers. In 4th Asia Conference on Intelligent Technology (ACIT). IEEE.
- [9] Omar, M., & Levkovich, I. (2024). Exploring the efficacy and potential of large language models for depression detection. *Journal of Affective Disorders*, 340, 1–12.
- [10] Assiri, A., Gumaei, A., & Mehmood, F. (2024). DeBERTa-GRU: Sentiment analysis for large language model. *Continuum: Lifelong Learning in Neurology*, Advance online publication.
- [11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171–4186.
- [12] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? *Proceedings of the China National Conference on Chinese Computational Linguistics*, 194–206.
- [13] He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding enhanced BERT with disentangled attention. *International Conference on Learning Representations (ICLR)*.
- [14] Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning (ICML 2003)*.
- [15] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610.
- [16] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
- [17] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259.
- [18] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610.
- [19] Neural Networks, 18(5–6), 602–610.
- [20] More, P.D., Sakhare, S.R., Joshi, A.U., Gaikwad, L.P., Jakite, V.P., Warbhuvan, A.S. (2023). Detection of Quality and Shelf Life of Fruits and Vegetables Using Neural Networks. In: Tuba, M., Akashe, S., Joshi, A. (eds) *ICT Infrastructure and Computing. ICT4SD 2023. Lecture Notes in Networks and Systems*, vol 754. Springer, Singapore. https://doi.org/10.1007/978-981-99-4932-8_53
- [21] A. Bhand et al., "DeepFish: A Comprehensive Exploration of Deep Learning Methods for Accurate Fish Species," 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2024, pp. 1-7, doi: 10.1109/ICCUBEA61740.2024.10774998.
- [22] P. More, A. Wanve, S. Shriram, S. Raskar and M. Sonawane, "Sports Video Summarization Using the Deep Learning Techniques," 2025 3rd International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 2025, pp. 1-6, doi: 10.1109/ICSSES64899.2025.11009492.
- [23] Singh, K., Mahalakshmi, V., Chavhan, P.G., Garg, P., Sahoo, G.S. and Sukumar, R., 2023, December. Investigating the Potential of Machine Learning for Enhancing Real-Time Data Analysis. In 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1-5). IEEE.
- [24] Chavhan, P.G., Chavhan, G.H., Shafi, P.M. and Mahalle, P.N., 2021. Context awareness computing system—a survey. In *Proceedings of International Conference on Communication and Artificial Intelligence: ICCAI 2020* (pp. 423-429). Springer Singapore