HSI-LFS-BERT: Novel Hybrid Swarm Intelligence Based Linguistics Feature Selection and Computational Intelligent Model for Alzheimer’s Prediction Using Audio Transcript

YUSERA FAROOQ KHAN1, BAIJNATH KAUSHIK2, MOHAMMAD KHALID IMAM RAHMANI2, (Senior Member, IEEE), AND MD. EZAZ AHMED2

1School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir 182320, India
2College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

Corresponding authors: Baijnath Kaushik (baijnath.kaushik@smvdu.ac.in) and Mohammad Khalid Imam Rahmani (m.rahmani@seu.edu.sa)

ABSTRACT Alzheimer’s dementia (AD) affects memory, language, and cognition and worsens over time. Therefore, it is critical to develop a reliable method for early detection of permanent brain atrophy and cognitive impairment. This study used clinical transcripts, a text-based adaptation of the original audio recordings of Alzheimer’s patients. This audio transcript data were taken from DementiaBank, which is the largest public dataset of AD transcripts. This study aims to show how Transfer Learning-based models and swarm intelligence optimization techniques can be used to predict Alzheimer’s disease. To enhance the prediction performance for Alzheimer’s disease, a hybrid swarm intelligence linguistic feature selection (HSI-LFS) approach is proposed that extracts a combined feature set using Particle Swarm Optimization (PSO), Dragonfly Optimization (DO), and Grey Wolf Optimization (GWO) algorithms. In addition, a transfer learning-based model called HSI-LFS-BERT, a combination of the HSI-LFS feature selection method and Bidirectional Encoder Representations from Transformer (BERT) algorithm, is proposed. The proposed model was compared using two feature sets: the first set consisted of the initial feature set and the second set contained a hybrid feature set that was extracted using the suggested HSI-LFS method. BERT embedding with HSI-LFS outperformed the conventional feature set, providing the most accurate modeling parameters while reducing the computations by 27.19%. The proposed HSI-LFS-BERT model outperformed state-of-the-art models, achieving 98.24% accuracy, 91.56% precision, and 98.78% recall.

INDEX TERMS HSI-LFS, machine learning, computational intelligence, particle swarm optimization (PSO), dragonfly optimization (DO), grey wolf optimization (GWO), Alzheimer’s dementia.

I. INTRODUCTION Alzheimer’s disease (AD) is one of the most common types of dementia worldwide. Alzheimer’s dementia (AD) is characterized by memory loss, which is an early symptom of the disease. The other symptom of AD is cognitive impairment, which affects the ability to make judgments and decisions, consequently causing difficulty with word retrieval. Alzheimer’s affects around 50 million people globally, and within the next 30 years, it is anticipated that more people may develop additional types of dementia due to the rise in the elderly population [1]. In India, the percentage has risen to 10.1% in 2021 and is expected to go up to 13.1% in 2031 [2]. AD affects one in every six people over the age of 80 and one in every 14 people over the age of 65 [3]. AD is neither curable nor reversible, but proper medication and...
can help it slow down or even stop its progression [4]. This has an impact on a patient’s day-to-day activities. They have the tendency of wandering and getting lost, taking longer than normal to complete daily tasks, repeating questions and so on. They have troubles such as involving in family gatherings, handling money, paying bills, engaging in sports, etc. The functional areas of their brain eventually degenerate, and they may lose complete control and become dependent on the permanent attention of a caregiver [5]. One of the most significant limitations of using learning models, including feature engineering, is the lack of availability of huge datasets. Multiple neuro-psychological examinations [6] involving cognition, concentration, executive function, linguistic, reasoning, and visuo perceptual abilities are carried out to acquire data for AD. For the diagnosis of AD, a list of methods and tasks used is shown in Fig 1. Verbal Fluency (VF), Spontaneous Speech (SS), and Other Tasks are the three main types of language tests used to determine symptomatic issues associated with AD [7]. Phonemic VF (PVF) and Semantics VF (SVF) are two types of VF tasks. In the PVF task, they were asked to think of as many words as they could, starting with the same letter (like the letter F) in one minute. In the SVF task, they are asked to name as many words from the same semantic category (like animal words) as they can in one minute.

FIGURE 1. Categories of linguistic test/exams for the diagnosis of AD.

The purpose of the SS activity is to ask the participants to describe a visual or to engage them in conversation. It could also be employed to recall a movie, a day of the week, a specific case, or a dream. In VF, success in exams is determined by an estimate of the number of correct words generated in one minute. The datasets related to AD contain audio or video recordings of these tasks, which have been transcribed and used in the literature. Using these datasets, researchers can study several aspects of cognition, semantic processing, and linguistic processes.

Audio transcript data of patients are investigated for AD prediction by using linguistic models that have been pre-trained to classify speech features as characteristics that are significant markers of the language and are dependent on the activities accomplished for diagnosis of the disease [8]. This not only addresses the issue of insufficient large datasets but also significantly minimizes the requirement for expert-defined features. The feature hybridization concept is introduced using a hybrid of three swarm intelligence (PSO, GWO, and DO) algorithms, demonstrating that it is a useful way of generating efficient prognostic models. An intelligent computational transfer learning model with hybrid feature selection approaches is proposed for AD prediction. The proposed model was also compared with various state-of-the-art models for performance evaluation.

A. MOTIVATION

Alzheimer’s speech assessment using Natural Language Processing (NLP) motivates this study. There is no cure for AD, but its progression can be delayed and, in some cases, halted through proper treatment if it is diagnosed early. The pathology of Alzheimer’s disease most likely starts several years before the symptoms appear [9]. As a result, there is a chance for prevention if potential developments make it possible to diagnose the disease using linguistic biomarkers before its symptoms appear. We analyzed written narratives to compare the language abilities of people with and without AD to investigate the connection between cognitive and language abilities and to find a potential predictor for its early detection. The motivation of the present work is to implement state-of-the-art techniques for automatic speech analysis to monitor AD patients and to shed light on prospective future research issues.

B. MAIN HIGHLIGHTS OF THIS WORK

The key contributions of this research work are summarized as follows:

- This study introduces the concept of feature hybridization using hybrid feature extraction techniques (HSI-LFS) inspired by swarm intelligence algorithms.
- The study proposed an HSI-LFS feature extraction method for disease classification using Transformer’s superficial backbone.
- The study also compared machine and deep learning backbones without word embedding, thus primarily emphasizing the feature extraction by the HSI-LFS architecture. It provides an extensive comparative study between feature extraction and word embedding techniques.
- The proposed model was compared using two feature sets: one containing word embeddings extracted using traditional techniques such as TF-IDF, Word2Vec, BERT, and XLNet, and the other containing word embeddings extracted using the suggested HSI-LFS method.
- The proposed architecture using HSI-LFS feature extraction along with BERT embedding provided the
most accurate modeling parameters and reduced computation by 27.19%.

The remaining part of this manuscript is organized as follows: Section II summarizes the related work of previous years. Section III discusses the dataset and presents a detailed methodology. Section IV presents the experimental details. Section V presents the results, performance analysis, and outcomes obtained through the hybrid features and the proposed model. Finally, the conclusion with future directions can be seen in Section VI.

II. RELATED WORK
Earlier works on language-based AD detection focused primarily on hand-crafted features extracted from transcripts with some acoustic data. Konig et al. collected data when participants completed a series of brief cognitive vocal tasks [10]. The vocal markers were detected and examined to discriminate between patients with Control Normal (CN), Mild Cognitive Impairment (MCI), and AD. The study distinguished CN and MCI by 79% ± 5% and that of CN and AD by 87% ± 3%. Orimaye et al. constructed ML models using the DementiaBank dataset for 99 patients with suspected AD and 99 others with CN, [11]. The model discovered a variety of syntactic, lexical, and n-gram linguistic biomarkers that could be used to identify differences between progressive AD and stable groups. Fraser and the team used the recordings of 264 people who talked about the Cookie Theft picture in the DementiaBank corpus. Automatic ML classification techniques have been used to differentiate between healthy and dementia patients with a standard accuracy of 81% [12].

Karlekar et al. [13] used three neural network models based on CNNs to identify AD and CN language samples. They explored the Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN). The accuracy of the classification models CNN, LSTM, CNN-LSTM, and CNN-LSTM with POS Tagged utterances were 82.8%, 83.7%, 84.9%, and 91.1%, respectively. In addition, Balagopalan, Aparna, et al. [14] applied feature-based ML and fine-tuned BERT classification model to analyze audio recordings and manually processed speech transcripts of 156 demographically older adults; 78 had AD, and 78 were healthy. The study reported 81.25% for the SVM and 83.32% for the BERT model for the classification of AD and healthy patients.

Xue et al. [6] worked on 1264 audio recordings of neuropsychological examinations administered to participants in the Framingham Heart Investigation (FHI). A two-level LSTM and CNN were applied to 483 subjects with CN, 451 with MCI, and 330 with dementia. The authors evaluated the results using 5-fold cross-validation and obtained a mean balanced accuracy of 0.647 ± 0.027 for LSTM and a mean area under the curve (AUC) of 0.805 ± 0.027 for CNN. Haulcy, R mani, and James Glass [15] experimented with the ADReSS challenge dataset that comprised 78 samples of non-AD and AD patients. The authors reported SVM and RF as top-performing models trained on BERT embeddings with audio and text features, obtaining 85.4% accuracy each. Classification accuracy in medical data for disease diagnosis can be improved by selecting an appropriate feature set and tuning the parameters of the classification model. Azadifar et al. [16] suggested a graph-theoretic gene selection for disease diagnosis. The method maximizes gene relevance to the target class and decreased inner redundancy by selecting the right genes from the maximal clique. Their model performed better than other filter-based gene selection techniques for cancer diagnosis. Saberi-Movahed et al. [17] introduced a Dual Regularized Unsupervised Feature Selection Based on Matrix Factorization and Minimum Redundancy (DR-FS-MFMR) method for choosing features. They experimented with nine gene expression datasets to demonstrate the performance of the DR-FS-MFMR method for improved feature selection. Performance evaluation showed that their proposed feature selection DR-FS-MFMR approach worked well and quickly for the feature selection task. Rostami et al. [18] proposed an innovative PSO-based multi-objective feature selection method that works in three steps. They tested the method on five medical datasets, and the results supported that their proposed method was more efficient and effective than the earlier methods. Rostami et al. [19] proposed a novel social network analysis-based gene selection method. The node centrality-based criterion was used to select the relevant and redundancy-free genes. Their technique improved the classification accuracy and time complexity. Khan et al. [20] suggested a Stacked Deep Dense Neural Network (SDDNN) model to predict AD by employing Glove and Randomly initialized embeddings with an accuracy of 93.31%. Based on the literature review, the authors identified these two research gaps: (1) It seems that there is a limitation in state-of-the-art architectures that consider multicollinearity and overfitting into consideration, and (2) There is a lack of availability of any transformer-based hybrid feature selection architecture. Consequently, we formulated two research questions to investigate further, as follows: RQ1: How can computation and multicollinearity from audio transcripts be reduced in determining Alzheimer’s Dementia? RQ2: How does the computational accuracy change within the influence of the proposed HSI-LFS-BERT?

III. METHODOLOGY
A model based on data-driven approaches may not be so promising for detecting AD using audio transcripts in the absence of a large dataset. Therefore, prior studies relied on features predefined by clinical experts, hence models capable of learning informative characteristics could not be utilized. Therefore, a framework consisting of highly pre-trained language models was used in this study to address this problem, as shown in Fig. 2.

A. DATASET DESCRIPTION
DementiaBank is a research corpus that collects speech and language samples of patients with AD and other
dementia-related disorders.\(^1\) This is one of the largest public datasets of transcripts and audio recordings from AD and normal patient interviews. The Pitt Corpus is a data collection of recordings of 264 participants who were asked to explain the image of Cookie Theft [21]. Cookie Theft is a sample image that participants are shown for its interpretation. The cookie Theft Image Test is a popular test for language evaluation and cognitive impairment. It is considered to be a common speech exercise for neurological diseases.

The dataset was partitioned into test and training data, and several classifiers were used to evaluate the effectiveness of automatic detection of Alzheimer’s dementia (AD+) and non-dementia (AD−) categories.

**B. DATA PREPROCESSING**

The data taken from DementaiBank was labeled as a .tsv file containing 3272 sample sentences. The sentences uttered by Alzheimer’s patients might have some grammatical or other language mistakes and would have repeated the same patterns in subsequent sentences. Therefore, data pre-processing is done as shown in Fig. 3.

The linguistic features identified in the dataset are converted into vectors using vectorization approaches. These features are listed as follows:

- **Short Answers and Bursts of Speech:** Formed clusters around short answers that have been split up by natural pauses in speech. For instance “uh”, “okay”, “s”, “um”, “hm”, “xxx”, “oh”, “shh”.
- **Repeated Clarification Requests:** Another formed cluster around questions for clarification and confusion about the task, specifically in the past tense. For instance “did I tell?”, “whether that’s more than what I said?”, “will he hit the bottom?”, “uh everything that’s going to happen, huh?”, “does that have enough?”, “I?”.
- **Starting with Interjections:** A lot of clusters contain utterances that start with interjections such as “oh”, “well”, “and”, “but”, “may be”. “maybe that was an apron and um maybe this was the um”, “oh there’s a girl &uh reachin(g) for a cookie”, “....

\(^1\)https://dementia.talkbank.org/access/English/Pitt.html
Based on these linguistic features, the data is labeled as 0 and 1; where 0 signifies normal patient (AD−) and 1 signifies an Alzheimer’s patient (AD+), as shown in Fig. 4 in which 1676 samples are from class 0 and the remaining 1596 samples are from class 1.

C. FEATURE EXTRACTION TECHNIQUES
This section describes the feature extraction techniques.

1) TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)
TF-IDF is a statistical method used for feature extraction [6]. It represents the frequency of a term (word) in a particular document if it appears in the document [22]. This approach captures the semantic meaning of words by identifying which words are important and which are irrelevant [23]. Each word in the collection has its TF and IDF value [24]. The fundamental premise driving TF-IDF is that a term’s significance is inversely correlated with its frequency across texts. A term’s frequency in a document is revealed by the TF, whereas IDF reveals its relative rarity among the collection of documents. The following are some of the important parameters:

- Stop words: This will be dropped from the actual collection of documents. Some stop words in English are: “is”, “an”, “this”, “the”, etc.
- n-gram range: by defining the minimum and maximum boundary, corresponding n-grams can be extracted.
- TF-IDF parameters: vector dimensions (i.e., number of documents in the entire corpus) = 316, maximum document frequency (max_df) = 0.9, and minimum document frequency (min_df) = 5. Here,
  - “max_df” is used to eliminate terms that appear very frequently. max_df = 0.9 signifies the exclusion of phrases that appear in more than ninety percent of the documents.
  - “min_df” is used to remove terms that occur infrequently. min_df = 5 signifies the exclusion of terms that occurred in fewer than 5 documents.

2) WORD TO VECTOR (Word2Vec)
Word2Vec is a traditional embedding that attempts to correlate a neighborhood of feature vectors using a Wikipedia dataset [25]. The embedding vector corresponds to a neighborhood kernel space of size d in a D-dimensional feature space. Embedding vector transposed d dimensional feature space correlating a neighborhood kernel space of size d. For selecting the kernel feature space, the word-to-vector embedding technique proposes two methods: a continuous bag of words (CBOW) or bag of words and n-grams. One benefit of the approach is that high-quality word embeddings can be learned efficiently with less time and space complexity.

3) BERT AND XLNet EMBEDDING
Transformer architectures contain a nemesis of attention-based mechanisms. Thus future extractions using multiple encoder and decoder functions are created. For transfer learning approaches, we utilized BERT embedding, which is the feature transformation of the objective function that allows decentralized encoder mapping of input and attention features [26]. For computational advantage, we used encoder layers in the BERT embedding followed by multiple perceptron layers to decode the feature extraction for disease classification. XLNet also uses similar architectural balances and the peculiar nature of tokenization, padding, and truncation with a specified function. The embedding uses 24 encoder layers and therefore creates a highly conjugated complex feature function for feature extraction and representation.

D. SWARM INTELLIGENCE FEATURE SELECTION TECHNIQUES
In this section, we describe the techniques for optimising feature selection.
1) PARTICLE SWARM OPTIMIZATION (PSO)
Kennedy and Eberhart developed PSO, a population-based stochastic searching approach [27]. The behavioural patterns of fish schooling, fish flocking, and insect swarms that fly across the search space in the best way to quickly approach the destination served as the inspiration for PSO. Each particle represents a potential solution in the D-dimensional search space. During the search process, data from each particle in the swarm is constantly evaluated and updated. The particles in this network navigate to a new location by modifying their velocity in accordance with both their expert judgment (personal best) and the collective intelligence of the swarm as a whole (global best).

2) GREY WOLF OPTIMIZATION (GWO)
Seyedali Mirjalili et al. developed GWO, a population-based meta-heuristics algorithm that replicates the way Grey wolves organize their leadership and hunt [28]. Metaheuristics have two key features: exploration and exploitation. Grey wolves have a social hierarchy and hunting method that includes three steps such as encircling the prey, attacking, and hunting the prey (exploitation) when the prey stops moving. Grey wolves finish hunting by attacking the prey and finally search for prey (exploration), where the grey wolves move away from the prey to find better prey.

3) DRAGONFLY OPTIMIZATION (DO)
DO was developed by Mirjalili Seyedali and features static and dynamic stages are equivalent to exploration and exploitation in metaheuristic optimization [29]. Dragonflies in the dynamic phase fly in larger swarms and in a single direction. Swarm consists of three fundamental phases: Separation, Alignment, and Cohesion.

a: THE FITNESS FUNCTION
calculates the canonical distribution of the log loss/cross entropy function based on cross-validated set results at every iteration of the genomic sequencing for the PSO, GWO, and GA architecture. The mathematical model for computing cross entropy and logarithmic loss are shown in Equations 1 and 2, respectively.

\[
\text{Entropy}(t) = \frac{1}{N} \sum_{j=1}^{M} [y_j \log(p(y_j))]
\]

\[
\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]
\]

where \(p(y_j)\) refers to the probability that a particular instance is considered in the summation loop. \(N\) represents the number of instances, \(y_i\) indicates the output of the \(i\)th instance and \(p_i\) represents the probability of 1, and \((1 - p_i)\) is the probability of 0. For the total number of cases, log loss accumulates the likelihood of a sample assuming both states 0 and 1. The distribution of feature space is such that the enhanced hybridized algorithmic structure uses the log loss validation curve as a generic fitness score and attempts to enhance the distribution using fitness optimization individually.

b: OBJECTIVE FUNCTION
The major purpose of feature selection is to reduce the fitness loss related to the respective swarm-based algorithms as shown in Equation 3).

\[
f(y_i, p_i) = \sum_{y_{(minimization)}} \left[ \left( \sum_{j=1}^{M} [y_j \log(p(y_j))] \right)
+ \left( -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \right) \right]
\]

The suggested architecture combines fitness functions to enhance the performance metrics and maximizes the performance accuracy of the model.

E. TRANSFER LEARNING MODELS
This subsection provides a brief description of the two transfer learning (TL) models implemented in this work for the classification of AD and non-AD patients.

1) XLNet
XLNet is an autoregressive language model trained using the Transformer-XL architecture. Zhilin Yang introduced the XLNet model and Generalized Autoregressive Pretraining for Language Understanding [30]. An autoregressive approach for all input sequence factorization orders was used to learn bidirectional contexts to improve bidirectional linguistic competence and word associations. The context word was used to forecast the following word. As a result, the subsequent token is dependent on all preceding tokens. The mask entropy function is reduced by XLNet, which was a feasible downside of the BERT initialization. XLNet is generalized because it utilizes a method called permutation language modeling to capture bi-directional context. We used embedding layers and dropout with a learning rate of 2e-5. The recurrence and attention concepts are combined in the transformer architecture to learn long-term dependencies. With state-of-the-art word embedding representation, it is an efficient model for classification.

2) BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)
The transformer system in BERT identifies contextual relationships between words in text [31] by using the decoder and the encoder part in order to make predictions based on the input data. Encoder-decoder system of the transformer leverages the attention mechanism to attain state-of-the-art effectiveness in the majority of NLP tasks [32]. The input embeddings were generated by adding a positional encoding of each word to a pre-trained embedding vector. To achieve
multi-head self-attention, Value (V), Key (K), and Query (Q) vectors are computed to extract different components of an input word. The feedforward layer is used in the encoder block to convert the output of the attention network to be used as input to the next encoder/decoder block. Masked Multi-head self attention generates attention vectors and therefore masking future words is important. Multi-Head Encoder-Decoder attention works with self attention vectors of the Masked Multi-head self attention and word attention vectors from the encoder block. BERT, is a state-of-the-art pre-trained language model that enables machines to learn enhanced representations of text compared to the context for natural language tasks.

IV. PROPOSED NOVEL APPROACH

In this study, we focused on joint feature selection by using swarm intelligence based methods and classification using pre-trained BERT and XLNet models. The proposed approach is executed in two phases: Phase 1 and Phase 2. In Phase 1, a hybrid feature set (X) is selected using the proposed novel hybrid feature selection approach. In Phase 2, two TL models (BERT and XLNet) we trained separately using a hybrid feature set (X) to make the final prediction, that is either AD+ or AD−.

A. PHASE 1: HYBRID SWARM INTELLIGENCE LINGUISTIC FEATURE SELECTION (HSI-LFS)

The feature selection approach attempts to find the best subset of features that is a key factor in predicting the disease. A model may be incapable of capturing the real behaviour of a patient with dementia if an unsuitable collection of features is selected. Therefore, the selection of relevant characteristics is a crucial part of model creation, as irrelevant features have a detrimental effect on the model performance. With a maximum length of 350 characters and 3724 samples in the dataset, 500-dimensional feature extraction allows for a lengthy computing operation. However, it is essential to note that brute force feature creation would face 2n-1 complexity, with n ranging from 300 to 400 feature vectors before embedding and 1024 vectors after embedding. As a result, the feature selection process would be time-consuming and challenging to maintain. To overcome these shortcomings, we propose a Hybrid Swarm Intelligence based Linguistic Feature Selection (HSI-LFS) approach to select the optimal features that would be provided to the proposed model for effective classification. This hybrid model of gated swarm intelligence algorithms includes Particle swarm optimization (PSO) [33], Grey wolf optimization (GWO) [34], and Dragonfly optimization (DO) [35] algorithms.

Optimal selection is necessary to decrease the computational complexity and obtain a uniform meaning from textual characteristics. The covariance in textual data is also reduced, which assists in the convergence of the gradient and fitness function to their optimal value as well as the loss caused by feature covariance. Algorithms 1, 2, and 3 demonstrate the formulation of the fitness functions for PSO, GWO, and DO respectively that are used in the proposed model. The proposed approach can be visualized as a master-slave

Algorithm 1 Fitness PSO

1: procedure ()
2: Def Log loss (Z, weights, total_weights, X) // Z is numpy.array containing training data with
3: shape(n_rows, n_cols) and X is numpy.array of coefficients with shape(n_cols)
4: scores = Z.dot(X)
5: Y = scores > 0
6: loss_value = np.empty_like(scores) // empty_like function creates a new array with the same shape and type as a
given array
7: loss_value[Y] = np.log1p(np.exp(-scores[Y]))
8: loss_value[~ Y] = - scores[~ Y] + np.log1p(np.exp(scores[~ Y]))
9: loss_value = loss_value.dot(weights) // weights is an NumPy.array of sample weights with
10: shape(n_rows,) and total_weights is NumPy.sum(total_weights)
11: return loss_value
12: end procedure

Algorithm 2 Fitness GWO

1: procedure ()
2: def cross_entropy_loss(yHat, y): // yHat represents
3: predicted values and y represents actual Values
4: if (y == 1):
5: return - np.log(yHat)
6: else:
7: return - np.log(1-yHat)
8: end procedure

Algorithm 3 Fitness Do

1: procedure ()
2: def cross_entropy_loss(yHat, y):
3: if (y == 1):
4: return - np.log(yHat)
5: else:
6: return - np.log(1-yHat)
7: end procedure
architecture using multi-lateral slave optimizers to demonstrate the loss and select features and master to accumulate the intersection of the most relevant features and calculate the final entropy. The step-wise functionality of this master-slave architecture is presented in Algorithm 4.

The intersection function allows feature vector combinations to convert feature space boundaries to the minimal parameterization of the results. Consequently, the feature extraction distribution is improved, and the bottleneck function can be further identified. The loss function allows the fitness vectors to improve the accuracy and other performance metrics and the intersection segment reduces the entropy and in turn fine tuning the model. It was found that the intersection of features decreased the entropy by more than 37% in comparison with the union and other metrics. To summarise:

- The log loss function is used as a genesis loss calculator to elaborate the fitness score for the individual architectures to run.
- The model uses master-slave optimization as a parental technique and genetic architectures as individual slaves.
- The final intersection results from the slave data forms were shared to be used by the master model.
- The Master model verifies the entropy space among the scope features of the union, intersection, and subtraction results.
- The model concludes that intersection is the optimal choice for comparison.

The input to the hybrid extractor is the full feature-set obtained by traditional methods represented by $\lambda$, where $\lambda = V_1, V_2, V_3, V_4, V_5 \ldots \ldots V_{n-2}, V_{n-1}, V_n$. PSO selected feature list is $\lambda_{PSO} = < V_1, V_3, \ldots V_j | j = 1, 2, \ldots, k >$, GWO selected feature list is $\lambda_{GWO} = < V_5, V_6, \ldots V_j | j = 1, 2, \ldots, k >$ and DO selected feature list is $\lambda_{DO} = < V_8, V_9, V_{11}, \ldots V_j | j = 1, 2, \ldots, k >$. The final feature set represented by $X$ is the concatenation of subsets produced by PSO, DO, and GWO to obtain multi-headed extraction views. The features extracted by utilizing the three optimization strategies substantially require less compilation time. The concatenation operation is demonstrated mathematically in Equation 4:

$$X = \sum \Pi_j^k \lambda(j, j : k)$$

where, $X = \text{Hybrid Features List (HSI-LFS)}$, $\sum = \text{Concatenation of } [\lambda_{PSO}, \lambda_{GWO}, \lambda_{DO}]$, $\Pi = \text{Feature Selector for each algo, } j = \text{index operator between list, } k = \text{final list index and } \lambda = \text{full feature list}$.

This hybrid feature selection approach (HSI-LFS) will be used in Phase 2 with BERT and XLNet model for classification of AD+ and AD−.

**B. PHASE 2: HYBRID CLASSIFICATION MODELS**

In phase 2, AD classification was performed using two models namely BERT and XLNet as discussed below.

1) HSI-LFS-BERT

HSI-LFS is used with BERT to create a hybrid model named HSI-LFS-BERT, which is used for predicting Alzheimer’s dementia. The detailed architecture of the Transformer BERT model is illustrated in Fig. 5.

The HSI-LFS-BERT architecture was trained where BERT was fine-tuned along with the hybrid feature set (X) for predicting AD+ and AD−.

The hybrid feature set was passed via the summary feature mapping to provide generic extraction via BERT. The BERT mechanism deforms audio transcripts into multi-headed sequence networks, which are then passed to various encoders demonstrating the relevance of each segment in the dataset. It carefully evaluates the masked transpositions and creates attention maps to emphasize Alzheimer’s textual contexts, allowing the upcoming network layers to utilize a more similar and non-deviated data sample space. PSO, GWO, and DO begins with a randomized initialization for every feature vector. However, with increasing iterations, the entropy decreases pertaining to the decrease in the loss scalar product generated by the y true and every scale model prediction. If the corresponding output feature subset decreases the entropy loss globally, the vector sample space is updated to the current feature space. This protocol uses the common

---

**Algorithm 4 Master-Slave for Feature Combination**

1: **procedure** ()
2: For each component until meeting the minimum or maximum iterations requirements
3: Create the particle
4: For every component
5: **Determine the fitness value**
6: Set current value as the new pBest if the fitness value is greater than the historically best fitness value (pBest)
7: End For
8: End procedure
9: **Particle position update**
10: Select the particle with the highest fitness value among all the particles to serve as the gBest for each particle.
11: **end procedure**

---

**FIGURE 5.** Demonstration of BERT training for audio transcript-based classification of AD+ and AD−.
objective function and utility of PSO, GWO, and DO, provides the gbest, X alpha, and X vectors, respectively. Based on the layered input, the concatenated unique feature list is passed to the BERT encoders, with six encoders and a pooling layer at the end, followed by a feed-forward neural network (FFNN) of shape 256. Internally, BERT encoders consist of multiple FFNN with shapes 512 and 1024. They are passed through a layer normalization and self-attention layer using a self-residual connection and bridge blocking method. The internal structure of each encoder is shown in Fig. 6.

![FIGURE 6. Demonstration of proposed model: HSI-LFS-BERT. The architecture of HSI-LFS-BERT consists of the proposed hybrid feature selection approach (HSI-LFS) and fine-tuned BERT model to make the prediction of AD+ and AD−.](image)

The output from each encoder is saved as per the encoder ID and passed to the pooling layer with 256 neurons. Pooling was used after the six encoders to determine the highest possible value for each patch in the feature map. Each layer's outputs are pooled into a single neuron in the max-pooling layer. The pooling layer independently processed each feature map to generate an equal number of pooled feature maps. The outcome of utilizing a pooling layer and producing pooled feature maps is a simplified version of the features observed in the input for making a prediction. TL models have brought breakthroughs in the field of language processing [36].

2) HSI-LFS-XLNet
The decoder of XLNet examines the full sentence to choose to extract information from it at any point in time because the order of tokens does not have to be comparable. As a result, the decoder allows access to all hidden states of the encoder. The decoder determines which hidden state to search by evaluating the hidden states of each encoder. A basic feed-forward neural network is used to determine the weights referred to as attention weights or values. XLNet embeddings were created in the study using a 24 encoder function vector, with a summarisation layer of 2148 vectors followed by five layered artificial neural networks, precisely 4096, 2048, 512, 128, and 2. With a decaying learning rate, the Optimizer RMSprop was utilized. The optimizer starts with a value of 0.083 and decays with the necessary dynamic momentum. The hidden state of the first segment is cached in memory by segment recurrence in each layer and attention is updated accordingly. Track of the position of each token in a sequence is done by positional encoding. It enables memory reuse for each segment. The categorical cross-entropy utilized in this architecture is the same as that used in the BERT architecture.

V. EXPERIMENTAL DETAILS
This section presents several experimental analyses of the proposed architectures to evaluate classification performance. The authors examined various performance criteria to evaluate AD prediction.

A. IMPLEMENTATION AND TRAINING DETAILS
The experiment was performed using a system with Intel’s i7 8th Generation processor with 32 GB RAM and a 16 GB Nvidia TX900 GPU. Tensorflow 3.1 and Keras were used as the backbones, providing the modeling libraries. In addition to the hardware requirements, Jupyter, Python, and the AWS console were used to implement the proposed models.

B. PARAMETER OPTIMIZATION
Out of the entire data set, 80% was used for training, and 20% was used as the test set. In classification models and estimators, parameters such as learning rate, constraints, batch size, and optimizer parameters are tuned during the training phase. The parameters provided before training are known as hyperparameters, and the optimal value selection process is known as hyperparameter tuning. To determine the optimal hyperparameters for the implemented models, we applied the Grid search technique [37]. This study considered an array of grid search parameters involving various levels of extraction and selection over the tree and other bagging architectures. The optimal parameter set selected and utilized for all the three approaches are listed in Table 1.

C. EVALUATION MATRICES
The confusion matrices [38] are required to provide an analysis of the number of written samples that were correctly classified as either a True Positive (TP), False Positive (FP), or True Negative (TN). We investigated five performance indicators: Accuracy, F1_Score, Recall, Area under the ROC curve (AUC) and confusion matrix for the three approaches applied to classification tasks in this study [39]. The results obtained using these five performance measures for binary class (AD+ and AD−) classification are provided in section VI.
TABLE 1. Optimized parameters of each model implemented in this study using Grid Search selection space optimization.

| Model   | Parameters                                                                 |
|---------|---------------------------------------------------------------------------|
| DT      | min_samples_leaf = 1, max_leaf_nodes = 1, max_depth = 5, min_samples_split = 2, n_estimators = 100 |
| RF      | n_estimators = 100, max_depth = 100, min_samples_split = 2, min_samples_leaf = 1 |
| SVM     | C = 1, kernel = linear, gamma = auto, max_iter = 1000, tol = 1e-6 |

TABLE 2. Comparative predictive analysis among three approaches with TF-IDF embedding and full feature list (lambda: \(\lambda\)) on DementiaBank dataset for the classification of AD+ and AD-.

| Model   | Testing Accuracy (%) | Validation Accuracy (%) | F1 Score |
|---------|----------------------|-------------------------|----------|
| Random Forest | 84.4                  | 85                      | 0.84     |
| Decision Tree  | 84.5                  | 85                      | 0.85     |
| Support Vector Machine | 85.0                 | 86                      | 0.85     |
| LSTM     | 85.4                  | 87                      | 0.84     |
| Bidirectional LSTM | 85.3                 | 87                      | 0.85     |
| CNN-LSTM | 90.0                  | 89                      | 0.90     |
| XLNet    | 92.0                  | 90                      | 0.92     |
| BERT     | 93.0                  | 95                      | 0.93     |

B. RESULTS WITH HYBRID FEATURE SELECTION APPROACH: HSI-LFS

For the chosen ML methods, Table 3 describes HSI-LFS feature selection strategies paired with TF-IDF word embedding. Table 4 shows deep learning methods such as LSTM, Bi-LSTM, and a hybrid of CNN-LSTM, along with transformer word embedding and HSI-LFS feature selection.

The rise in accuracy and other metrics from Table 3 to 4 was considerable. The Random forest architecture achieves 84.4 percent accuracy, whereas HSI-LFS feature selection obtains 90.53 percent accuracy. A similar improvement may be seen in a decision tree architecture, which has gone up from 84.5 percent to 89.61 percent in accuracy. However, there is a decrease in the accuracy of the SVM due to inefficient kernel optimization. The SVM does not synchronize internally with the master-slave architecture proposed in this study. This is because the SVM kernel auto-transposes the results and undergoes feature selection which is reduced to 68.39% in our approach. Hence it shows a decrease in synchronization leading to a decrease in accuracy.

Table 5 shows a comparison of the typical complete feature set and HSI-LFS set with appropriate transformer embedding. Between Table 3 and 5, there is a difference of 3-5% in the accuracy metric. Without feature selection, the BERT architecture achieves an accuracy of 91 percent, an F1_score of 90.17 percent, and a specificity of 94.5 percent, although there is a feature selection change of 5 to 9% overall in all the accurate measurements. A similar finding was reached With XLNet, with delta changes ranging from 1 to 6%. The proposed HSI-LFS-BERT model achieved a benchmark accuracy of 98.74%. The performance comparison of the proposed approach to state-of-the-art approaches with different embedding methods and HSI-LFS is shown in Fig. 7. The plot depicts that BERT with HSI-LFS showed better performance for AD+ and AD− classification task.

C. COMPUTATIONAL ANALYSIS BETWEEN TRADITIONAL AND HYBRID FEATURE SELECTION

Table 6 compares the computational time of the algorithm for full feature extraction without embedding, feature extraction with embedding, and feature selection with the HSI-LFS hybrid feature extractor.

The proposed framework works in the de-entanglement of collinear features from dimensional biases reducing the probability of overfitting and allowing a robust unbiased modeling system. The PSO reduces GPU cycles by 13.77%, DO 12.74% GPU cycles, and GWO 17.79% GPU cycles. The proposed HSI-LFS generates an optimized feature selection channel that reduces time complexity by 27.19% while maintaining the accuracy and other metrics of the state-of-the-art standards discussed above.

D. CONFUSION MATRIX

The four prediction outcomes that occurred while performing the classification of AD are shown in Fig 8 using the
The proposed HSI-LFS-BERT model showed better performance by predicting 159 TP and 145 TN values, whereas HSI-LFS-XLNet predicted 144 TP and 135 TN values. In addition, HSI-LFS-BERT predicted 7 FP (AD+) and 17 FN (AD−), whereas HSI-LFS-XLNet predicted 9 FP (AD+) and 20 FN (AD−). This clearly shows that the proposed HSI-LFS-BERT model achieves a better prediction performance for AD prediction.

**E. PLOTS FOR PERFORMANCE VISUALIZATION**

We plotted the training and validation accuracies across the number of epochs in the training process for

| Algorithm  | With lambda (λ) | With Embeddings | With HSI-LFS (X) |
|------------|-----------------|----------------|-----------------|
| DT         | 155             | 208 > 416 > 420 | 141             |
| RF         | 214             | 260 > 500 > 501 | 214             |
| SVM        | 64              | 124 > 350 > 355 | 62              |
| LSTM       | 2489            | 2497 > 2714 > 2710 | 2444            |
| Bi-LSTM    | 2585            | 2498 > 2700 > 2709 | 2200            |
| CNN-LSTM   | 2857            | 3014 > 3219 > 3220 | 2814            |
| HSI-LFS-XLNet | 5547         | 5569 > 5600 > 5650 | 5327            |
| HSI-LFS-BERT | 5048       | 5104 > 5400 > 5410 | 5000            |
HSI-LFS-XLNet, as shown in Fig. 9. This will help in interpreting important decisions regarding the model’s performance.

The training loss versus validation loss over epochs for HSI-LFS-BERT is depicted through graphs shown in Fig. 10. BERT has shown higher validation accuracy and lower validation loss than HSI-LFS-XLNet. As an illustration, the corresponding ROC curves [40] for the ML, DL, and TL techniques were drawn in Fig 11, 12 and 13 respectively.

To find the best algorithm for the binary classification of AD+ and AD− patients, AUC scores were compared and the model with a higher AUC score was preferred. The inference drawn is that the area under the entire curve is used to measure diagnostic accuracy. The AUC values for XLNet and HSI-LFS-BERT are 0.95 and 0.97, respectively.

The ROC curve and AUC for HSI-LFS-BERT demonstrate that it is a better predictor of AD in healthy individuals.

F. CROSS-VALIDATION (CV)

Furthermore, Cross-validation (CV), a statistical technique, was implemented in this work to generalize the applicability of models in terms of performance. We performed 10-fold cross-validation (K=10), where the data set was divided into ten parts at random.

Nine of them were used for training, and one-tenth were used for testing and the process was repeated ten times with a fresh holdout set and a new tenth to be tested. For k-1 training iterations, every data point is subject to an exact test. The variance in the estimate diminishes as k increases. Moreover, CV has a higher potential for enabling the detection of overfitting. Table 7 lists the performance of the eight models tested in this study for K=10 folds, and the mean for each model was calculated over the 10 folds. The statistical mean following cross-validation illustrates the performance of the model over ten subgroups of the entire sampling. On average, 89.32% of the ML models, such as DT, exhibit a strong comprehension and feature simulation. A larger distributed mean for BERT (95.52%) than for XLNet (90.97%) and CNN-LSTM hybrid architecture (90.80%) indicates better stability of the BERT model. Furthermore, the standard deviation (StdDev) is a measure used to evaluate the variation in the scores. When performing cross-validation for models, the standard deviation of the scores of k folds is also evaluated and is listed in Table 7. The models show a relatively
stagnant comparison of the standard deviation and average accuracy values however, XLNet is the most stable model with a standard deviation of 0.00117.

### G. COMPARISON OF THE PROPOSED METHOD WITH OTHER ADVANCED TRANSFORMER BACKBONES

Table 8 illustrates the superiority of the proposed method over other encoder-decoder based transformer architectures such as Roberta, Distilbert, and XLNet.

The proposed method outperformed the other three algorithms with more than 98.77% accuracy, 98.707% F1_score, 98.709% precision, and 98.84% recall.
VII. CONCLUSION
This paper presents state-of-the-art natural language processing models in the field of healthcare research. Alzheimer’s dementia (AD) can be diagnosed more quickly and accurately if linguistic biomarkers are analyzed through the verbal utterances of elderly persons, which is evident from experimental evaluations. We implemented the TL approach, which performs efficiently in the classification of linguistic data and compared it with conventional ML and DL models. Statistical ML classifiers such as SVM and RF treat the sentence as a bag of word models (each word is treated independently irrespective of its position in the sequence). Bi-LSTM architectures are ace in capturing the bidirectional semantic context of words in each sentence. They can create near-realistic language models and thereby providing a better classification of the sentences. HSI-LFS, a hybrid swarm intelligence based feature selection method proposed in this study with a combination of Particle swarm optimization (PSO), Dragonfly optimization Algorithm (DO), and Grey wolf optimization algorithm (GWO), aids RQ1 in reducing multicollinearity and overfitting criterion from the dataset. The HSI-LFS uses deep learning techniques with gradient boosting and adaptive height updates to train a hierarchical model capable of detecting and classifying high-level labels in any two-dimensional space. The Hybrid feature selection (HSI-LFS) enhanced the performance of learning models by reducing competition and boosting performance. Among the two transformer models, BERT and XLNet employed with HSI-LFS, the proposed novel HSI-LFS-BERT method obtained benchmark accuracy of 98.74%, the precision of 98.83%, and F1_Score of 98.826% for prediction of AD+, thereby aiding the RQ2 for proposing a state-of-the-art model reducing variance and multicollinearity but maintaining the computational accuracy. In the future, experiments should be performed to determine which biomarker is linked to which brain pathology. Neurodegenerative diseases, such as Alzheimer’s, Parkinson’s, and Amyotrophic Lateral Sclerosis are being investigated to determine how they differ from each other. Precise diagnosis from speech or signal data and these multimodal evaluations may assist in distinguishing between various types of brain pathologies.

DATA AVAILABILITY
1. The python code of this article is available on GitHub at the URL:
   https://github.com/yusrafarooq/HSI-LFS-BERT_ORG
2. The data underlying this article is available at URL: https://dementia.talkbank.org/access/English/Pitt.html

FUNDING STATEMENT
The authors received no specific funding for this study.

CONFLICTS OF INTEREST
The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES
[1] World Health Organization. WHO’s Work on the Global Dementia Observatory. Accessed: Jul. 32, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia
[2] Business today. Accessed: Jun. 6, 2022. [Online]. Available: https://www.business today.in/lifestyle/health/story/indias-elderly-population-grows-to-136-cr-in-2021-over-lower-death-rates-nso-study-304209-2021-08-15
[3] National Health Service. About Dementia. Accessed: Jul. 17, 2022. [Online]. Available: https://www.nhs.uk/conditions/dementia/about/
[4] G. E. Stutzmann, “The pathogenesis of Alzheimers disease—Is it a lifelong ‘calciumopathy’?” Neuroscientist, vol. 13, no. 5, pp. 546–559, Oct. 2007.
[5] C. Tunnard, D. Whitehead, C. Hurt, L. O. Wahlund, P. Mecocci, M. Tonali, B. Vellas, C. Spencer, I. Kloszewksa, H. Soninen, S. Lovestone, and A. Simmons, “Apathy and cortical atrophy in Alzheimer’s disease,” Int. J. Geriatric Psychiatry, vol. 26, no. 7, p. 741–748, 2011.
[6] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, and V. B. Kolachalam, “Detection of dementia on voice recordings using deep learning: A Framingham heart study,” Alzheimer’s Res. Therapy, vol. 13, no. 1, pp. 1–15, Dec. 2020.
[7] F. Galaverna, A. M. Bueno, C. A. Morra, M. Roca, and T. Torralva, “Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia,” Eur. J. Psychiatry, vol. 30, no. 4, pp. 305–320, 2016.
[8] L. Brabenec, J. Mekyska, Z. Galaz, and I. Rektorova, “Speech disorders in Parkinson’s disease: Early diagnostics and effects of medication and brain stimulation,” J. Neural Transm., vol. 124, no. 3, pp. 303–334, Mar. 2017.
[9] E. McDade and R. J. Bateman, “Stop Alzheimer’s before it starts,” Nature, vol. 547, no. 7662, pp. 153–155, 2017.
[10] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, “Automatic speech analysis for the assessment of patients with predementia and Alzheimer’s disease,” Alzheimers’ Disease, Diagnosis, Assessment Disease Monit., vol. 1, no. 1, pp. 112–124, Mar. 2015.
[11] S. O. Orimaye, J. S.-M. Wong, K. J. Golden, C. P. Wong, and I. N. Soiyiri, “Predicting probable Alzheimer’s disease using linguistic deficits and biomarkers,” BMC Bioinf., vol. 18, no. 1, pp. 1–13, Dec. 2017.
[12] K. Fraser, J. Meltzer, and F. Rudzicz, “Linguistic features identify Alzheimer’s disease in narrative speech,” J. Alzheimer’s Disease, vol. 49, no. 2, pp. 407–422, 2016.
[13] S. Karlekar, T. Niu, and M. Bansal, “Detecting linguistic characteristics of Alzheimer’s dementia by interpreting neural models,” 2018, arXiv:1804.06440.
[14] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, “Comparing pre-trained and feature-based models for prediction of Alzheimer’s disease based on speech,” Frontiers Aging Neurosci., vol. 13, Apr. 2021, Art. no. 635945.
[15] R. Hauley and J. Glass, “Classifying Alzheimer’s disease using audio and text-based representations of speech,” Frontiers Psychol., vol. 11, Jan. 2021, Art. no. 624137.
[16] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, and M. Oussalah, “Graph-based relevancy-redundancy gene selection method for cancer diagnosis,” Comput. Biol. Med., vol. 147, Aug. 2022, Art. no. 105766.
[17] F. Saberi-Movahed, M. Rostami, K. Berahmand, S. Karami, P. Tiwari, M. Oussalah, and S. S. Band, “Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection,” Knowl.-Based Syst., vol. 256, Nov. 2022, Art. no. 109864.
[18] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, “Integration of multi-objective PSO based feature selection and node centrality for medical datasets,” Genomics, vol. 112, no. 6, pp. 4370–4384, Nov. 2020.
[19] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, “Gene selection for microarray data classification via multi-objective graph theoretic-based method,” Artif. Intell. Med., vol. 123, Jan. 2022, Art. no. 102228.
[20] Y. F. Khan, B. Kaushik, M. K. I. Rahmani, and M. E. Ahmed, “Stacked deep neural network model to predict Alzheimer’s dementia using audio transcript data,” IEEE Access, vol. 10, pp. 32750–32765, 2022.
[21] P. Klump, J. Fritsch, and E. Nöth, “Ann-based Alzheimer’s disease classification from bag of words,” Proc. 13th ITG-Symp. Speech Commun., Frankfurt, Germany: VDE, 2018, pp. 1–4.
[22] V. Kumar and B. Subba, “A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus,” in Proc. Nat. Conf. Commun. (NCC), Feb. 2020, pp. 1–6.
Learning, neuroimage image analysis, and computer vision.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–11.

S. A. Shah and Y. Y. Ou, “TRP-BERT: Discrimination of transient pattern recognition, machine learning, and deep learning.

B. Subba and P. Gupta, “A TfidfVectorizer and singular value decomposition based host intrusion detection system framework for detecting anomalous system processes,” Comput. Secur., vol. 100, 2021, Art. no. 102084.

R. B. Shamantaha, S. M. Shetty and P. Rai, “Sentiment analysis using machine learning classifiers: Evaluation of performance,” in Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS), 2019, pp. 21–25, doi: 10.1109/CCCOMS.2019.8821650.

R. Wang and Y. Shi, “Research on application of article recommendation algorithm based on Word2Vec and TFIDF,” in Proc. IEEE Int. Conf. Electr. Eng., Big Data Algorithms (EEDBA), Feb. 2022, pp. 454–457.

M. Liu, L. Liu, J. Cao, and Q. Du, “CO-attention network with label embedding for text classification,” Neurocomputing, vol. 471, pp. 61–69, Jan. 2022.

J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proc. Int. Conf. Neural Netw. (ICNN), vol. 4, 1995, pp. 1942–1948.

S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey wolf optimizer,” Adv. Eng. Softw., vol. 69, pp. 46–61, Mar. 2014.

S. Mirjalili, “Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems,” Neural Comput. Appl., vol. 27, no. 4, pp. 1053–1073, 2016.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized autoregressive pretraining for language understanding,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–11.

S. M. A. Shah and Y.-Y. Ou, “TRP-BERT: Discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT,” Comput. Biol. Med., vol. 137, Oct. 2021, Art. no. 104821.

A. Özyiçift, K. Akarsu, F. Yumuk, and C. Şoytemez, “Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): An empirical case study for Turkish,” Automatika, vol. 62, no. 2, pp. 226–238, Apr. 2021.

K.-L. Du and M. Swamy, “Particle swarm optimization,” in Search and Optimization by Metaheuristics, Springer, 2016, pp. 153–173.

A. Zareie, A. Sheikhhahmadi, and M. Jalili, “Identification of influential users in social network using gray wolf optimization algorithm,” Expert Syst. Appl., vol. 142, Mar. 2020, Art. no. 112971.

C. M. Rahman and T. A. Rashid, “Dragonfly algorithm and its applications in applied science survey,” Comput. Intell. Neurosci., vol. 2019, pp. 1–21, Dec. 2019.

S. Ruder, M. E. Peters, S. Swayamsiddha, and T. Wolf, “Transfer learning in natural language processing,” Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Tuts., 2019, pp. 15–18.

P. C. Bhat, H. B. Prosper, S. Sekmen, and C. Stewart, “Optimizing event selection with the random grid search,” Comput. Phys. Commun., vol. 228, pp. 245–257, Jul. 2018.

A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” Pattern Recognit., vol. 91, pp. 216–231, Oct. 2019.

M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, Jul. 2009.

K. Hajian-Tilaki, “Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation,” Caspian J. Internal Med., vol. 4, no. 2, p. 627, 2013.