# JASBO: Jaya Average Subtraction Based Optimization with Deep Learning Model for Multi-Classification of Infectious Disease from Unstructured Data

*Vian Sabeeh** 🆔✉ *Ahmed Bahaaulddin A. Alwahhab* 🆔✉ *Ali Abdulmunim Ibrahim Al-kharaz* 🆔✉

Informatics Department, Technical College of Management-Baghdad, Middle Technical University, Baghdad Iraq.
*Corresponding Author.

## Abstract

Infectious diseases have become an unavoidable big trouble in today's environment with a similar symptomatology that makes difficult of early detection and clear separation of infection. Hence, it is required to generate a new technique that best utilizes the various symptomatologies present in the illnesses for its multi-classification. Medical documents are considered an essential source for modern, invented, and robust analysis methods for accurate infection diagnoses. Accordingly, enriching medical text processing is beneficial in health informatics. In this research, proposed Jaya Average Subtraction Based Optimization (JASBO), which is enabled by Deep Learning (DL) is used to classify infectious diseases into many categories from unstructured data. Moreover, the DL model used is Infectious Disease Network (ID-Net) which combines Convolutional Neural Network (CNN) and Bidirectional-Long Short-Term Memory (Bi-LSTM). To specify the strange or discriminative words with BI-LSTM . JASBO algorithm used in the model to determine the size of the filter in the final classification network to detect the meaningful part of the text. The input text is given to the Tokenization layer in this case, where the tokens get formed and is forwarded to CNN. Additionally, character-based network features are extracted using Bi-LSTM model. Then, vector representation is concatenated with two separate character-level extractions from Bi-LSTM and CNN. Character level features are passed to the attention layer, which uses the Kumar-Hassebrook similarity measure to calculate the score function. Label of each word token is then predicted by the ID layer, at which layer size is found by JASBO. Here, JASBO combines Jaya algorithm with an Average and Subtraction-Based Optimizer (ASBO). The best performance of JASBO_ID-Net is analyzed with three performance metrics: accuracy with superior value of 91%, recall with high value of 88.7%, and F-measure with a superior value of 90%.

**Keywords:** Average and Subtraction-Based Optimizer (ASBO), Bidirectional-LSTM (Bi-LSTM), Convolutional Neural Network (CNN), Infectious Disease Network (ID-Net), Jaya algorithm.

## Introduction

The availability of medical data has increased recently, leading to more extensive usage of Electronic Health Records (EHR) in hospital settings. The various medical and healthcare tasks, such as population healthcare and other clinical decision support, rely on the massive amount of data [1]. EHRs, which are digital databases of patient medical incidents as well as observations, are already

pervasive in medicine that is essential for operations, administration, and research. According to the formats used for collection and depiction, data in EHRs is frequently divided into structured and unstructured data. The medications, diagnoses, and laboratory terms seen in structured EHR data are organized into set numerical or categorical fields. Contrarily, free-form language generated by health care professionals, like discharge summaries and clinical notes, is referred to as unstructured data. About 80% of all EHR data is unstructured, but it is still exceedingly challenging to handle this data for later use [2]. As EHR systems have been adopted so quickly, the majority of patient data is now kept in an electronic format. Due to the data's dispersion across numerous systems and formats, necessary data for retrospective research investigations is challenging to gather [3]. While being a significant data source for health research, EHRs often contain unstructured data that can be challenging to extract. Due to its ease of analysis, coded data has been employed in most research up to this point, although clinical entries' unstructured "free" text may also include crucial information. Manually reviewing free text takes time, and anonymization may be necessary to safeguard patient privacy [4].

A scenario that results from the invasion of the human body by a harmful agent that injures the body and may spread to people is known as an infectious disease. Infectious illnesses are the main cause of sickness and mortality for any given population. Infectious diseases include infections with the Human Immunodeficiency Virus, Acquired Immune Deficiency Syndrome (HIV/AIDS), Hemagglutinin type 1 and Neuraminidase type 1 (H1N1) influenza, Poliomyelitis, and Severe Acute Respiratory Syndrome (SARS). Although the prevalence of infectious diseases varies widely between nations on a global scale, in the twenty-first century, hepatitis B, hand-foot-mouth disease, and tuberculosis are most common. These diseases significantly impact the health and economies of the entire world. Infectious diseases continue to pose a serious threat to the health of a population despite multiple vital advances in the methods for avoiding and controlling different diseases. The spread of infectious diseases is influenced by factors such as population growth, environmental changes, and rising antibiotic resistance. 57 million deaths reported annually around the world were caused by infectious diseases, accounting for more than 25% of all the fatalities [5]. Infections that appear at the present or appeared in

the past are spreading quickly in a wide geographical area and may be specified as emerging infections. Emerging diseases have altered the trajectory of human history and brought about untold suffering and death [6]. Newly emerging infectious illnesses significantly impact global economies and public health. Although it is believed that socioeconomic, environmental, and ecological factors have a significant role in their genesis. There is no systematic comparative research that examined these relationships to distinguish the temporal and spatial infections patterns of disease [7].

One of the most life-threatening and dangerous diseases affecting individuals worldwide is infectious diseases; They continue to be a major global basis of mortality, morbidity, disability, as well as socioeconomic instability despite medical advancements. Lower respiratory infections, which account for 20% of all deaths and are the main cause of mortality overall among infectious diseases, are primarily causing 90% of pneumonia-related deaths among children under five. Sepsis is a significant clinical issue that develops in hemodialysis patients with advanced chronic kidney disease, along with respiratory infections that can be bacterial, viral, or fungal in origin [8]. Making proper suspected infectious diseases diagnosis is crucial for the prevention as well as control of infectious diseases while dealing with a variety of illnesses. Artificial Intelligence (AI) is currently being utilized to classify infectious diseases. The detection of lymph node metastases in breast cancer patients, the dermatological level classification of skin cancer, the diagnosis of different forms of Alzheimer's disease, diabetic retinopathy, and diabetic macular edema all make use of AI techniques [9]. Many have used DL algorithms to predict disease in recent years due to their improved performance. Data elements in EHR to be utilized, data representations to get developed for feeding into DL models, and design of learning models all impact how well the prediction performs when using EHR data to build DL models. Because EHR data is readily available, multiple DL models are utilized to various EHR data aspects for predicting disease. For instance, using EHR data, Long Short Term Memory (LSTM) network model can predict both diabetes and heart failure [10]. Despite recent advancements in infectious disease detection and classification, this is still difficult to detect and classify automatically. Moreover, multi-classification of infectious disease is a long-term process as the symptomatologies remain same.

Hence, it is necessary to find a proper method for multi-classifying infectious diseases from unstructured data.

This research depends on the multi-classification of infectious diseases from unstructured data. First, a dataset's input text sentence is obtained, and then ID-Net is used to classify infectious diseases into multiple categories. Text input is provided to the tokenization layer in this ID-Net, a new DL architecture built on a combination of Bi-LSTM and CNN. Tokens fed towards CNN are further allowed towards Bi-LSTM, at which character-level features are obtained. CNN and Bi-LSTM are thus used for extracting character-level information from networks. Next, a vector representation is concatenated with two separate character-level extractions from Bi-LSTM and CNN. Then, character-level features are supplied to attention layer, which calculates the score function using the Kumar-Hassebrook similarity metric. The label of each word token is then predicted by ID layer, which is above the attention layer. Here, JASBO is an amalgamation of ASBO and Jaya algorithm that is used to calculate the layer size of ID layer.

The main contribution of this paper is developing JASBO_ID-Net for the multi-classification of infectious diseases. A multi-classification of infectious disease from unstructured data is gained by using ID-Net at which layer size is found by JASBO. Here, JASBO is formed by combining both Jaya algorithm and ASBO. This combination is very helpful in solving real-world problems like disease classification.

The remaining work of this is involved with section two enumerates related work; section three provides JASBO_ID-Net details for multi-classification of infectious disease from unstructured data. Section four brings about results and discussions; and the last section is the conclusions.

## Related Work

Wang M. et al. [9] used the Multiple Infectious Disease Diagnostic Model (MIDDM) for multi-classifying infectious diseases from unstructured electronic medical records. The proposed model used an auto-encoder to deal with the sparsity problem of data to improve the training set for the model. The auto-encoder does not need to use the infectious disease category to which the sample belongs as the label but learns the characteristics of the sample as the input of the neural network and the label of the model concurrently. Each document type is entered into the auto-encoder for feature extraction as a vector. Then the feature vector is fed into softmax for classification. This model proved very important for earlier detection and infectious illness warning. However, this approach was not demonstrated to be effective for the early detection of rare infectious illnesses. Vidhya, K. and Shanmugalakshmi, R. [11] suggested Deep Belief Network (DBN) for diabetes complication prediction. The model controlled diabetes well in contrast to Type 2 Diabetes (T2D) condition and can successfully handle large amounts of medical data generated by healthcare facilities and extended its assistance to society by assisting in the advanced prediction of complex diseases. Despite these advantages, this model has not been extended to handle additional types of diabetes as well as its inability to take advantage of a cloud-based patient monitoring system for greater accuracy. Wang, S.M., et al. [12] presented the International Statistical Classification of Disease and Related Health Problems 10th revision code –Classification Model (ICD-10 CM) that primarily used free-text medical notes to automatically derive the corresponding diagnosis codes. The model is a neural network that consists of a word embedding layer and a bidirectional Gated Recurrent Unit (GRU) layer that outputs the prediction. However, a rule-based classification system was not created for the ICD-10-CM's delicate regulations, such as combination codes. Compared to a professional programmer, this model significantly reduced the personnel required for coding. However, this strategy required a lot of data for training, and having more data led to problems with data imbalance. Zhao, J., et al. [13] designed Multimodal Deep Feature Fusion with next-generation sequencing technology for auxiliary diagnosis model of infectious respiratory diseases. This method had low training time and was quick and accurate. However, this approach only resolves sequence data. Maheshwari, V., et al. [14] utilized a Fuzzy-based Decision Tree (FDT) algorithm to predict Coronavirus Disease 2019 (COVID-19). The paradigm of the model relies on building fuzzy classification rules from the dataset and optimizing the gained rules by using a genetic algorithm according to the accuracy fitness target. Medical data structures and incomplete data, this technology reconstructs lost or incomplete data because of the ability of the model to build fuzzy rules from the dataset itself. However, this approach did not

successfully strike a balance between data privacy and public health when using AI interactions. The FDT algorithm made it easier to spot nondeterministic occurrences in unstructured datasets pertaining to medical diagnosis. Luo, X., et al. [10] schemed ensemble models like DL and Machine Learning (ML) to identify chronic cough patients using EHR data. With this technique, it was simple to locate chronic cough without using the International Statistical Classification of Disease (ICD) code. Further illness prediction, however, was not found in this model.

Venkataraman, G.R., et al. [15] suggested a model for automatic text classification of unstructured medical narratives. Medical narratives were automatically categorized with very little human preprocessing. The model maps each word into a specific numeric vector relying on the Global vector of words (Glove). The Recurrent Neural Network (RNN) model trained on features extracted from meta map processing includes nouns derived from each word and other derivates. Regardless of the data provided, this approach was not demonstrated to categorize every clinical notion appropriately. Nagamine, T., et al. [16] presented a model of unsupervised clustering medical records to predict multi-scale heart failure. It used an automatic pattern recognition of real-world disease symptoms to understand complex syndromes, but it was unable to track the progression of heart failure patients' disease states over time. This method served as foundation for practice-based medicine; However, more complaints were taken from clinical text in EHR. Although there are numerous approaches for early detection and treatment of infectious diseases, this prediction remains challenging due to the lack of available data and human error.

## Materials and Methods

Diagnosing infectious diseases helps in providing a better treatment and enhances the prevention and controlling diseases. The proposed model consists of two parts. The first part is the ID-Net which utilizes deep learning for multi-classification of infectious disease from the unstructured data, While the second part is the JASBO which is the combination of ASBO [20] and Jaya algorithm [21] used to determine the layer size of ID-Net to predict the label of infectious diseases. Fig.1 depicts the general diagram of JASBO_ID-Net for multi-classification of infectious disease.

The advancement of disease diagnosis has witnessed significant evolution through the integration of modern technologies, particularly the Internet of Things (IoT) and its application in healthcare settings. IoT for healthcare has introduced novel methods for diagnosing diseases, including infectious diseases. Several researchers explored the potential of IoT-enabled healthcare solutions to enhance diagnostic processes. Ahmed A. et al. [17] addressed the challenges of energy consumption and security in healthcare systems by proposing an IoT healthcare secure energy saver protocol for medical applications. Their research focused on optimizing energy usage while ensuring the security of medical data. Additionally, their work introduced a transmission protocol beneficial for patient monitoring, which can have implications for real-time disease classification. While S. Ashraf et al. [18] contributed to the field by presenting an efficient node coverage system applicable in healthcare environments. They highlighted the utility of sensors for transmitting medical documents, indicating the potential for IoT-based data exchange in disease diagnosis. Their algorithm for sensor placement aimed at achieving maximum coverage through a distributed movement approach, ultimately reducing computation time and increasing the effectiveness of the system. Furthermore, the work of S. Ashraf et al. [19] expanded beyond the confines of healthcare diagnostics. They delved into efficient supply chain management strategies and environmentally friendly product and service design. Recognizing the impact of optimal supply chain management on healthcare delivery efficiency, their research indirectly contributes to the broader healthcare landscape, potentially influencing disease diagnosis processes.

Initially, an input text sentence is acquired from the dataset [22], which is then fed to the first part, the ID-Net, for the multi-classification of infectious diseases. This ID-Net is the new DL architecture based on the integration of CNN [23,24] and Bi-LSTM [25], where text input is given to the tokenization layer. The obtained tokens are then fed to CNN, where character-level features are acquired. Those features are allowed to Bi-LSTM, employed to extract network character-level features. The dual various character level representations taken from Bi-LSTM and CNN are concatenated with vector

representation. Then, character level features are fed to the attention layer that calculates the score function using the Kumar-Hassebrook similarity measure [26], and input to the ID layer above the attention layer predicts the label of every word token. Here, the role of part two, represented by JASBO, is taken to determine the layer size to predict the infectious disease.
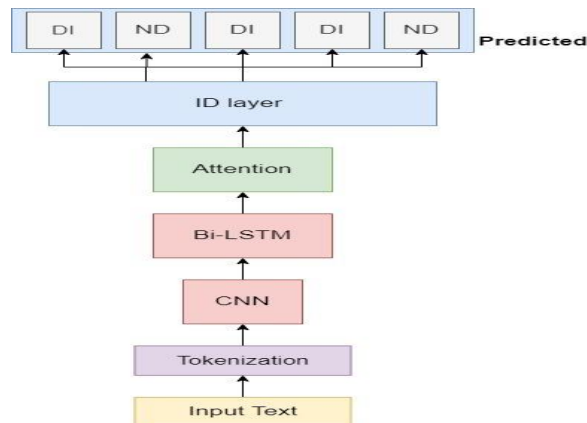


**Figure 1. The general diagram of JASBO_ID-Net**

**ID-Net Architecture**

ID-Net is the first part of the newly developed framework in this work for the multi-classification of infectious diseases from unstructured data. Text is taken as input towards ID-Net that is applied to the tokenization layer at which input sentence is changed to words or tokens. Based on the tokens formed, a matrix is generated based on zero and non-zero padding. Then, the generated matrix is forwarded to CNN, having multiple layers like the convolutional layer, batch normalization, Rectified Linear Unit (ReLU), maximum pooling layer, Fully Connected (FC) layer, and softmax layer. At CNN, feature vectors are generated. CNN has the advantage of extracting high-level features by using convolutional and max-pooling layers but has a shortage in capturing longer dependencies and sequential patterns. Therefore, the output of the CNN layer is fed toward Bi-LSTM to extract low-level character patterns from a text in order to gain more temporal relationships and contextual understanding. In Bi-LSTM, there are layers like embedding, Bi-LSTM, and dropout, at which character-level features are generated. Further, the two different character level features are concatenated at the FC layer, and then at the attention layer, the tagging score is generated. From the tagging score, a label is predicted for each token by the ID layer. In Bi-LSTM, there are layers

like embedding, Bi-LSTM, and dropout, at which character-level features are generated. Further, the two different character level features are concatenated at the FC layer, and then at the attention layer, the tagging score is generated. From the tagging score, a label is predicted for each token by the ID layer.

**Tokenization Layer**

The tokenization layer is the initial layer at which the input text from the dataset is forwarded. In the tokenization layer, the words or tokens are formed from the input sentence. The number of tokens depends on the word numbers present in the input sentence.

**Matrix Formation**

After the conversion of sentences into tokens, the matrix is formed based on the letter or character of words. The matrix formed is filled with padding values and zero padding values. If the character is present, it resembles padding value 1, and if character is absent at that position, it is represented as zero. Thus, matrix is formed from the character of tokens and is indicated as $M$. The matrix formed is shown $M$ with size $(l \times p \times m)$, which is given as input to CNN.

**CNN**

The matrix $M$ formed is sent to CNN to extract not only coarse features but also more complex and abstract features that gradually lead to fine features; Therefore three CNNs has been used depending on the continues experimental and validation test. CNN consists of various layers like convolution layer, batch normalization, ReLU layer, maximum pooling layer, FC layer, and softmax. In CNN, character-level features are extracted and the various layers of CNN are explained below,

**Convolution Layer**

The most significant layer in CNN is convolution layer, which has many successive layers that result from layer is sent to successive layers as input. Moreover, consecutive layers are responsible for the accuracy of the method. The input is first fed to this layer, which forms trust as well as nodes energy and from those features, more confined features get extracted. Neurons present in successive convolution layers are connected by trainable weight sets via receptive fields. Here, a nonlinear activation function

is created between successive convolution layers and input. In this layer, input is given as $M$ and the output formed is represented as in Eq.1,

$$Y = \{Y_1, Y_2, \ldots, Y_f, \ldots Y_g\}$$

1

where, $g$ indicates total convolutional layers that act on the input and find the result as in Eq.2,

$$\left(X_f^\ell\right)_{u,v} = \left(E_f^\ell\right)_{u,v} + \sum_{i=1}^{h_i^{f-1}} \sum_{r=-\chi_1^f}^{\chi_1^f} \sum_{s=-\chi_2^f}^{\chi_2^f} \left(v_{\ell,i}^f\right)_{r,s} * \left(X_f^{\ell-1}\right)_{u+r,v+s}$$

2

where a convolutional operation for extracting local patterns from result of successive convolutional layers is *, the convolutional layer's fixed feature map $f$ is specified as $\left(X_f^\ell\right)_{u,v}$, the feature map relating to the previous convolutional layer is $\left(X_f^{\ell-1}\right)_{u+r,v+s}$, the kernel function in $f^{th}$ convolution layers is indicated as $v_{\ell,i}^f$, and bias is represented as $E_f^\ell$. The convolutional layer indicates the size of $(l \times p \times m)$.

### Batch Normalization

Batch normalization [27] is the next layer, followed by the convolutional layer. The input distribution is transformed through batch normalization into a normal distribution. This serves to speed up training as well as using high learning rates to make learning easy. This conversion causes the distribution to fall within the activation function's sensitive interval, which means a minor change in the input results in a significant loss function change. Additionally, by making the gradient bigger, the issue of gradient dispersion is avoided. $(l \times p \times m)$ is the size indicated for the batch normalization layer.

### ReLU Layer

ReLU is an activation function that removes negative values for effectiveness and simplicity. In ReLU, operations like division, multiplication, or exponential operations are not considered. Hence, the computational operation of ReLU is more straightforward. As every network layer indicates nonlinearity, ReLU is idempotent. Here, $f^{th}$ the

nonlinear layer along feature maps and resultant from ReLU is indicated by Eq. 3,

$$X_f^\ell = fun\left(X_i^{\ell-1}\right)$$

3

where $fun(\ )$ is the activation function in the convolution layer $f$. ReLU dimension is indicated by size $(l \times p \times m)$.

### Maximum Pooling Layer

The pooling layer helps in reducing data dimensions by joining neuron cluster outputs at one layer to a single neuron in the next layer. The maximum pooling layer is noticed in this architecture as this acts as sub-sampling layer via minimizing feature map resolutions for enhancing features' invariance to distortions available in the input data. This layer has no bias or trainable weight as it is non-parameterized. $(l \times p \times m)$ is size designated for the maximum pooling layer.

### FC Layer

In FC layer, every neuron joins to all neurons of the previous layer, which makes this layer generate many parameters in this layer. This generation of parameters helps as a prediction module of characters in words. This layer acts as secure nodes and the result from the FC layer is indicated by Eq. 4,

$$Q_f^\ell = \varepsilon(H_f^\ell)\, with\, H_f^\ell =$$
$$\sum_{i=1}^{h_i^{f-1}} \sum_{r=-\chi_1^f}^{\chi_1^f} \sum_{s=-\chi_2^f}^{\chi_2^f} \left(v_{\ell,i}^f\right)_{r,s} * \left(X_f^{\ell-1}\right)_{u+r,v+s}$$

4

where, $\left(v_{\ell,i}^f\right)_{r,s}$ is weight linking unit at $(u,v)$ in $i^{th}$ feature map of $(f-1)$ layer and $\ell^{th}$ unit in $f$ layer. FC layer is expressed with dimension $(l \times c)$.

### Softmax Layer

The softmax layer [27] is always followed by the FC layer. In the softmax layer, the elements are rescaled to form a particular function. $(l \times d)$ is the size represented for the softmax layer. Various CNNs connected so that character-level features get extracted from them. The output from CNN is

represented by term $C$, and extracted features are then allowed to the Bi-LSTM layer.

## Bi-LSTM

The output from CNN that is indicated as $C$ are then fed to Bi-LSTM for extracting character-level features. Bi-LSTM considers both backward and forward information for capturing hidden states of dual LSTMs. The layers of Bi-LSTM are the embedding layer, Bi-LSTM, dropout layer, as well as FC layer.

## Embedding Layer

The embedding layer [28] in Bi-LSTM is characterized by word and character embedding. Word embedding is mapping words to lower-dimensional vectors. The merit of utilizing a word vector is that this captures the semantics of words or relationships among them. Along with word embedding, character-level embedding also indicates input tokens. Character-level embedding is helpful for effectively extracting morphological information from every word token. The dimension of an embedding layer is represented by $(l \times n \times d)$.

## Bi-LSTM Layer

Bi-LSTM is employed to extract character-level features. Bi-LSTM is applied over character embedding sequence for every word and dual final hidden states from both backward and forward LSTM [29] are concatenated for obtaining a fixed-size vector, which represents a word token. By utilizing bidirectional hidden states, the model preserves future and past information. Moreover, Bi-LSTM captures the global features of every word token effectively. $(l \times d)$ is the size of the Bi-LSTM layer.

## Dropout Layer

In the dropout layer, the dimension size is dropped out, where the dimension size is reduced. This dropout is applied to the final hidden state of the Bi-LSTM layer, and the size is indicated as $(l \times d)$. The output from the dropout layer of Bi-LSTM is represented by $D$.

## FC Layer

FC layer is employed for connecting the character features of both Bi-LSTM and CNN. Next, a vector representation is concatenated with two separate character-level extractions from CNN and Bi-LSTM. They help to connect the features and are fed to attention layer. FC layers output is indicated by $F$ and $(l \times c)$ is the size of FC layer.

## Attention Layer

Attention layer [25,28] helps in focusing certain features, and this approach takes the attention of both CNN features and Bi-LSTM features. The output from the FC layer $F$ is allowed to the attention layer. The attention technique is used to find relations effectively among words that enable effective recognition of entities. The attention mechanism helps in finding score function by capturing correlation among words in the sentence using the Kumar-Hassebrook similarity measure, which is indicated by Eq. 5,

$$KH = \frac{\sum(R*S)}{\sum R^2 + \sum S^2 - \sum(R*S)} \quad 5$$

where, $KH$ is the Kumar-Hassebrook similarity measure $R$ and $S$ indicates the character level features for which score is obtained. Attention layer output is fed to ID layer. The attention layer result is represented as $L$, which is further allowed to ID layer. Size of the attention layer is indicated as $(l \times 2)$.

## ID Layer

ID layer takes attention layer output $L$, and at ID layer, each word token's label is predicted based on tagging score. In ID layer, the appropriate word is identified and the label is predicted at which the infectious disease word "thrombolytic" is identified. As multiple layers are present in ID layer, the layer size is determined for ID layer with the help of JASBO. The output gained from the ID layer is indicated as $P$.

## Parameters to Be Optimally Tuned

For finding the layer size of the ID layer, a newly developed optimization algorithm, namely JASBO, is proposed. Here, JASBO is formed by combining both Jaya algorithm and ASBO. Based on JASBO, the kernel size of the convolutional layer, batch normalization, ReLU layer, and maximum pooling is

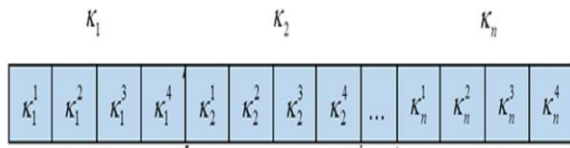determined. Fig. 2 represents the solution encoding of JASBO.



**Figure 2. Solution encoding of JASBO**

Here, $\kappa_1$, $\kappa_2$, and $\kappa_n$ indicates kernel size of CNN layer 1, 2, and $n$. Moreover, $\kappa_1^1$, $\kappa_1^2$, $\kappa_1^3$ and $\kappa_1^4$ are kernel size of convolutional layer-1, batch normalization-1, ReLU layer-1, and maximum pooling-1. Thus, the solution is encoded by JASBO, and parameters are optimally tuned.

### JASBO for Determining Layer Size of ID Layer

JASBO is the newly developed optimization algorithm formed by combining both the Jaya algorithm and ASBO. In ASBO, average information and subtraction of best as well as worst population members are found to guide algorithm population in problem search space. Here, population members transferring in search space to aim at reaching a suitable quasi-optimum solution. Also, Jaya algorithm is a powerful optimization algorithm to solve unconstrained and constrained optimization issues. Jaya algorithm depends on the concept that the solution gained for a given issue moves towards the ideal solution and avoids the worst solution. This algorithm needs regular control parameters and doesn't need any specific control parameters. Thus, combined algorithm, JASBO is very supportive in resolving problems in an easy way and avoids any inappropriate consequences. The procedure indicating JASBO is given as followed,

### Step 1: Initialization

In ASBO, every member population is the solution to the optimization issue. ASBO is a vector with element numbers that equal decision variable numbers. In ASBO, every element indicates value of corresponding variable element. This population of ASBO is indicated in vector form as in Eq. 6,

$$J = \begin{bmatrix} J_1 \\ \vdots \\ J_t \\ \vdots \\ J_W \end{bmatrix}_{W \times y} = \begin{bmatrix} j_{1,1} & \cdots & j_{1,x} & \cdots & j_{1,y} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ j_{t,1} & \cdots & j_{t,x} & \cdots & j_{t,y} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ j_{W,1} & \cdots & j_{W,x} & \cdots & j_{W,y} \end{bmatrix}_{W \times y}$$

6

where candidate solution is indicated as $J$, $J_t$ is $t^{th}$ candidate solution of ASBO, $y$ is a number of decision variables of the problem, $j_{1,x}$ is the value of $x^{th}$ decision variables found by $t^{th}$ candidate solution, and $W$ is members in ASBO. Each member of ASBO enables the objective function, which is indicated by Eq. 7,

$$E = \begin{bmatrix} E_1 \\ \vdots \\ E_t \\ \vdots \\ E_W \end{bmatrix}_{W \times 1}$$

7

where, $E_t$ indicates objective function value corresponding $t^{th}$ member, and $E$ is objective function vector.

### Step 2: Evaluating Fitness

Fitness evaluation is used in finding the best optimal solution for every object gaining appropriate outcomes. This value is gained from the result of ID-Net as well as targeted output that is indicated using Eq.8,

$$F_{ness} = \frac{1}{\vartheta} \sum_{\rho=1}^{\vartheta} \left[ Tar_p - P \right]^2$$

8

where, $F_{ness}$ is fitness value, $\vartheta$ is processing samples of full count, $P$ is result from ID-Net, and $Tar_p$ is targeted output.

### Step 3: First Stage of ASBO

In this stage, the average of the worst and best members of the population is attained using Eq.9 as,

$$X^{q1} = \frac{J_z + J_o}{2} \qquad 9$$

Here, $X^{q1}$ is the average of the best and worst population members, $J_o$ is the worst member of ASBO, and $J_z$ is the best member of ASBO, and Eq. 10 to Eq.12 represent the calculation of $j_{t,x}$.

$$j_{t,x}(U+1) = j_{t,x}(U) + e.\left(X_x - T.j_{t,x}(U)\right) \qquad 10$$

$$j_{t,x}(U+1) = j_{t,x}(U) + e.X_x - e.T.j_{t,x}(U) \qquad 11$$

$$j_{t,x}(U+1) = j_{t,x}(U)[1 - e.T] + e.X_x \qquad 12$$

where, $X_x$ is $x^{th}$ the dimension of $X^{q1}$, $T$ is a random number that equals 1 or 2, and $e$ is a random number that ranges $[0,1]$.

The basic formula of the Jaya algorithm is represented in Eq. 13 as,

$$j_{t,x}(U+1) = j_{t,x}(U) + e_1$$
$$\left(j_{t,x,best}(U) - \left|j_{t,x}(U)\right|\right) - e_2\left(j_{t,x,worst}(U) - \left|j_{t,x}(U)\right|\right) \qquad 13$$

where, $e_1$ and $e_2$ are random numbers.

Assume, $j_{t,x}(U)$ as positive, Eq.14 to Eq.16 forms the value of $j_{t,x}(U)$

$$j_{t,x}(U+1) = j_{t,x}(U) + e_1$$
$$\left(j_{t,x,best}(U) - j_{t,x}(U)\right) - e_2\left(j_{t,x,worst}(U) - j_{t,x}(U)\right) \qquad 14$$

$$j_{t,x}(U+1) = j_{t,x}(U)[1 - e_1 + e_2]$$
$$+ e_1 j_{t,x,best}(U) - e_2 j_{t,x,worst}(U) \qquad 15$$

$$j_{t,x}(U) = \frac{j_{t,x}(U+1) - e_1 j_{t,x,best}(U) + e_2 j_{t,x,worst}(U)}{[1 - e_1 + e_2]} \qquad 16$$

Substitute Eq. (16) in Eq. (12),

$$j_{t,x}(U+1) = \left[\frac{j_{t,x}(U+1) - e_1 j_{t,x,best}(U) + e_2 j_{t,x,worst}(U)}{[1 - e_1 + e_2]}\right] \qquad 17$$
$$[1 - e.T] + e.X_x$$

$$\frac{j_{t,x}(U+1)[1 - e_1 + e_2] - j_{t,x}(U+1)[1 - e.T]}{[1 - e_1 + e_2]} =$$
$$\left[\frac{\left(e_2 j_{t,x,worst}(U) - e_1 j_{t,x,best}(U)\right)[1 - e.T] + e.X_x[1 - e_1 + e_2]}{[1 - e_1 + e_2]}\right] \qquad 18$$

$$j_{t,x}(U+1)[1 - e_1 + e_2 - 1 + e.T] = \left(e_2 j_{t,x,worst}(U) - e_1 j_{t,x,best}(U)\right)$$
$$[1 - e.T] + e.X_x[1 - e_1 + e_2] \qquad 19$$

$$j_{t,x}(U+1) = \frac{1}{eT - e_1 + e_2} \qquad 20$$
$$\left[\left(e_2 j_{t,x,worst}(U) - e_1 j_{t,x,best}(U)\right)[1 - e.T] + e.X_x[1 - e_1 + e_2]\right]$$

Eq.17 to Eq.20 forms the basic equation of JASBO.

### Step 4: Second Stage of ASBO

In this phase, the population position is updated by subtraction information of worst and best population members. This is simulated by Eq. 21 to Eq. 23,

$$X^{q2} = J_z + J_o \qquad 21$$

$$j_{t,x}^{new,q2} = j_{t,x} + e.X_x^{q2} \qquad 22$$

$$J_t = \begin{cases} J_t^{new,q2}, & E_t^{new,q2} < E_t \\ J_t, & else \end{cases} \qquad 23$$

where, $X^{q2}$ is the subtraction of best and worst members of ASBO, $J_t^{new,q2}$ is the novel value of $t^{th}$ candidate solution based on stage 2, $E_t$ is its objective function, as well as $j_{t,x}^{new,q2}$ is $x^{th}$ the dimension of $J_t^{new,q2}$.

### Step 5: Third Stage of ASBO

In the third phase, the best member is used to lead the population of ASBO to the best solutions, which is simulated using Eq.24 and Eq.25,

$$j_{t,x}^{new,q3} = j_{t,x} + e.\left(j_{t,x} - T.j_{z,x}\right) \qquad 24$$

$$J_t = \begin{cases} J_t^{new,q3}, & E_t^{new,q3} < E_t \\ J_t, & else \end{cases} \qquad 25$$

where, $J_t^{new,q3}$ is novel status of $t^{th}$ population member based on stage 3, $E_t^{new,q3}$ is its objective function, and $j_{t,x}^{new,q3}$ is $x^{th}$ a dimension of $J_t^{new,q3}$.

### Step 6:End

JASBO is ended depending on reevaluation of fitness as in Eq. 8. Thus, while the maximum solution is attained, the iteration process gets end. This brings the best solution of JASBO for the multi-classification of infectious diseases. Algorithm. 1 indicates the pseudo-code of JASBO.

Algorithm 1. JASBO

| |
|---|
| **Input:** Search agent $W$, iterations $U$, objective function $E$, and constraints. |
| **Output:** Optimum solution $j_{t,x}(U+1)$ |
| Step 1 : **Start** |
| Step 2 : Initializing population by Eq. (6) |
| Step 3 : Find fitness by Eq. (8) |
| Step 4 : Evaluate objective function |
| Step 5 : For $U$ as iteration |
| Step 6 : Update worst and best members of population |
| Step 7 : For $t = 1\,to\,W$ |
| Step 8 : Phase 1: |
| Step 9 : Calculate $X^{q1}$ by Eq. (9) |
| Step 10 : **Hybridization of Jaya algorithm with ASBO** |
| Step 11 : Find basic equation of JASBO by Eq. (20) |
| Step 12 : Phase 2: |
| Step 13 : Calculate $X^{q2}$ by Eq. (21) |
| Step 14 : $J_t$ is updated by Eq. (23) |
| Step 15 : Phase 3: |
| Step 16 : $J_t$ is updated by Eq. (25) |
| Step 17 : end |
| Step 18 : Re-evaluate by fitness as per Eq. (8) |
| Step 19 : Find best solution |
| Step 20 : **End JASBO** |

### Method Implementation

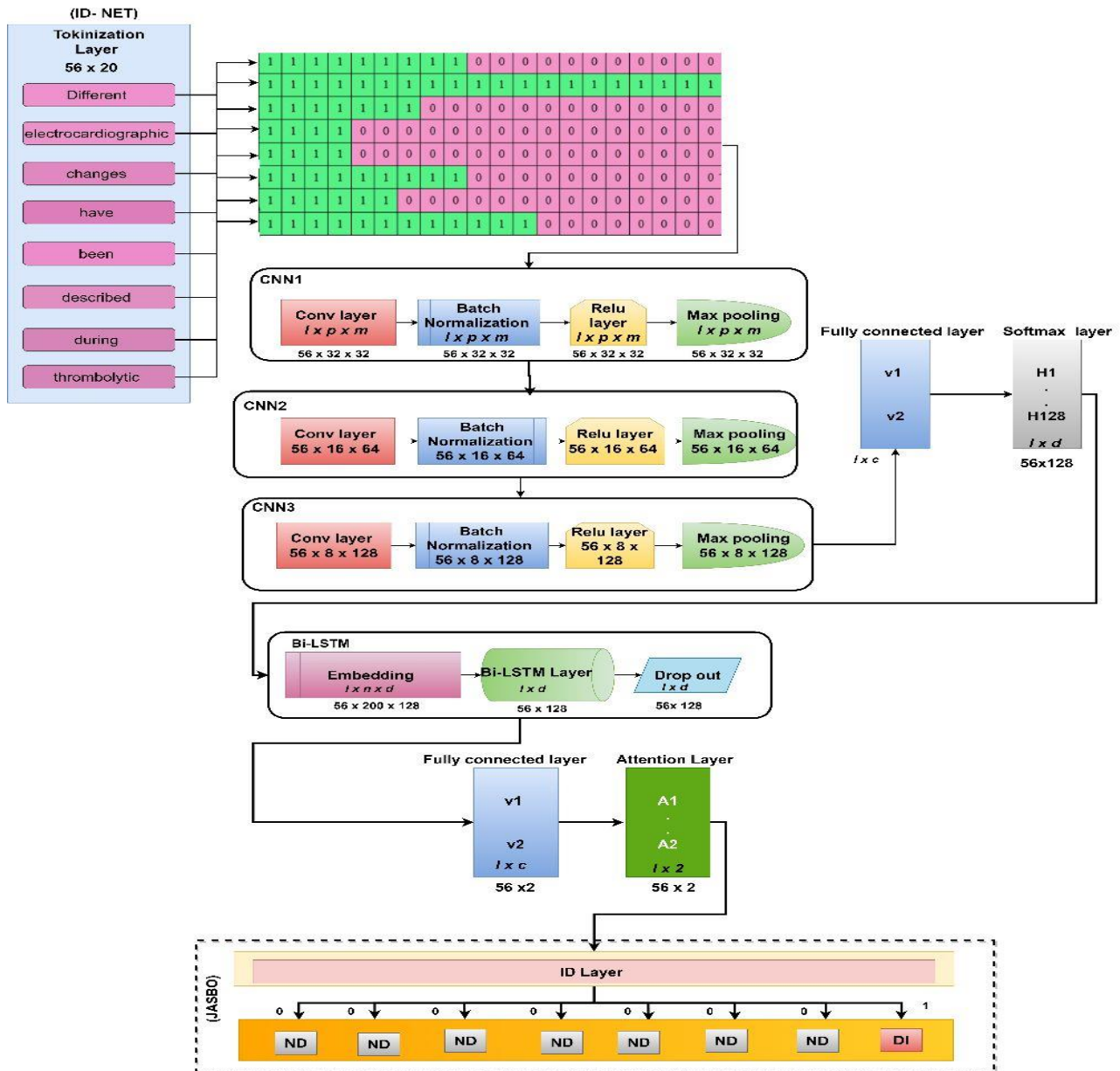Fig.3 illustrates the implementation details of JASB_ID-NET.

**Figure 3. JASBO_ID-NET Implementation**

Initially, the input sentence "Different electrocardiographic changes have been described during thrombolytic" acquired from a dataset is allowed to ID-Net having tokenization as initial phase. In tokenization, the above sentence is converted into various tokens or words such as, "Different," "electrocardiographic," "changes," "have," "been," "described," "during" and "thrombolytic". These words are arranged into zero padding and padding matrix based on letters in words. This matrix has values of 0 and 1 with dimension $(56 \times 20)$. This is then moved to CNN, with various layers like convolution, batch normalization, ReLU, and maximum pooling layers.

Three sets of CNN are used in ID-Net with first CNN layers of sizes $(56 \times 32 \times 32)$, second CNN layers of sizes $(56 \times 16 \times 64)$, and third CNN layers of sizes $(56 \times 8 \times 128)$. Here, in CNN, the character level features get extracted and this is then fed to FC layer $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$ with dimension $(56 \times 2)$ Then, this is forwarded towards softmax layer $\begin{bmatrix} H_1 .....H_{128} \end{bmatrix}$ with size $(56 \times 128)$. Further, for extracting character level features, Bi-LSTM is utilized, having layers of embedding with dimension $(56 \times 200 \times 128)$, Bi-

LSTM layer with dimension $(56 \times 128)$, and dropout with dimension $(56 \times 128)$. Then, the generated features are concatenated and moved to the FC layer $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$ with dimension $(56 \times 2)$ and the attention layer $\begin{bmatrix} A_1 & A_2 \end{bmatrix}$ with dimension $(56 \times 2)$. Finally, the ID layer utilizes up the score value generated in

## Results and Discussion

JASBO_ID-Net is analyzed with various performance measures and the results are depicted in this section.

### Experimental Setup

The experimental setup of this paper is done in a Python ver 3.11 programming language with libraries such as numpy [30], keras [31], and sklearn [32].

### Dataset Description

The Medical Dataset for Abbreviation Disambiguation for Natural Language Understanding (MeDAL) dataset is used for conducting this research. It has been carefully extracted from Public Medline (PubMed) search engine that contains about 18,374,626 medical abstracts and released during the EMNLP ClinicalNLP workshop in 2019 with 14,393,619 medical articles. It is a sizeable medical text dataset curated for abbreviation disambiguation and designed for natural language understanding pre-training in the medical domain. The existence of abbreviations with their descriptor text and their locations in the text makes this dataset more beneficial. It can be used with various applications like information retrieval, question answering, clinical decision support, medical document summarization besides developing and evaluating Natural Language Processing (NLP) models. The (MeDAL) dataset was initially used for a training model that can solve medical terms disambiguated by the training model depending on the text context. While in this paper, this dataset is used for diagnosing infectious diseases through the abbreviations mentioned within the context of medical text. The dataset was labeled using the reverse substitution paradigm without human labeling. This way is conducted by specifying full idioms in the text which have known abbreviations,

attention layer to access the words or tokens to end up with multi-classification of infectious disease from the text. Value 1 resembles that the disease is present, indicated as DISEASE (DI) while value 0 resembles that the disease is not present, indicated as NO DISEASE (ND). Here, it is found that the word "thrombolytic" is the identified disease from the input text.

then replacing them with their meaningful abbreviations [33].

### Evaluation Metrics

In this paper, three metrics like accuracy, recall, and F-measure are used that are explained below,

### Accuracy

This indicates an accurate prediction of disease from overall infections. This brings the performance of the model across every class and is generated normally by the ratio of accurate infectious predictions to overall infectious predictions, expressed by Eq.26,

$$A_y = \frac{t(+) + t(-)}{t(+) + t(-) + f(+) + f(-)}$$

26

Here, $A_y$ is accuracy, $t(-)$ is true negative, $f(-)$ is false negative, $f(+)$ is false positive, and $t(+)$ is true positive.

### Recall

This is a measure that brings out infectious prediction count as positive as well as original. This resembles best fraction indicating true positive to overall count of true positive and false negative that is noted using Eq.27,

$$R_l = \frac{t(+)}{t(+) + f(-)}$$

27

Here, $f(-)$ is false negative, and $R_l$ is recall.

### F-measure

This is created from recall with precision and calculated for finding average detection rate of infectious disease and is formulated by Eq.28,

$$F_e = \frac{t(+)}{t(+) + \frac{1}{2}(f(+) + f(-))}$$

28

Here, $f(+)$ is false positive, and $F_e$ is noted for F1-score.

**Performance Assessment**

Figure 4 is a performance assessment of JASBO_ID-Net with three metrics like accuracy, recall, and F-measure with varying epoch sizes.


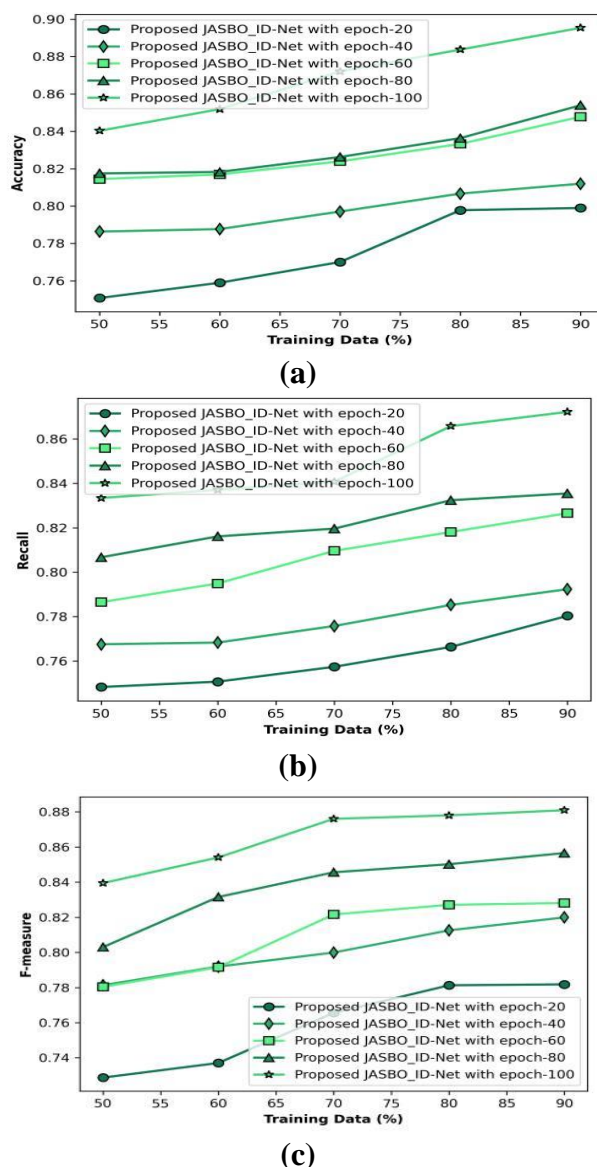
**(a)**



**(b)**



**(c)**

**Figure 4. Performance analysis of JASBO_ID-Net, a) accuracy, b) recall, and c) F-measure**

Fig.4 (a) resembles accuracy-based performance assessment. When training percentage is 50%,

accuracy is 0.840 for JASBO_ID-Net with epoch 100, wherein accuracy is attained as 0.751, 0.786, 0.814, and 0.817 for JASBO_ID-Net with epoch 20, 40, 60, as well as 80. Fig.4 b) is recall based performance assessment of JASBO_ID-Net. While 60% is taken as training data, recall is 0.751, 0.768, 0.795, 0.816, and 0.837 for JASBO_ID-Net with epochs 20, 40, 60, 80, and 100. Fig.4 c) indicates F-measure based performance assessment. While 90 is considered as training data, then F-measure is high for JASBO_ID-Net with epoch 100 of 0.881, whereas JASBO_ID-Net with epoch 20, 40, 60, as well as 80, attains an F-measure of 0.782, 0.820, 0.828, and 0.857.

**Comparative Methods**

Medical data like the Health Insurance Portability and Accountability Act (HIPAA) in the United States is sensitive. Revealing such data may be considered a strict privacy regulation. Sharing sensitive medical data is challenging, as they often contain personally identifiable information that may need to be secured. Therefore, collecting datasets that previous research papers have used is very difficult unless publicly available. The proposed model must undergo an efficiency test with a fair comparison. So, the tests must be implemented on various previous models on the MeDAL dataset. The various comparative models used in this paper to compare with JASBO_ID-Net are DBN.[11], ICD-10 CM[12], FDT algorithm[14], and MIDDM[9].

**Comparative Assessment**

The comparison of JASBO_ID-Net is done based on varying training percentages and k-value, as given below,

**Comparative Evaluation Using Different Training Percentages**

Fig.5 indicates a comparison by varying training percentages. Fig.5 (a) denotes accuracy-related comparative assessment by altering training data. When training percentage is 90%, then accuracy is 0.839, 0.861, 0.882, 0.878, and 0.902 for DBN, ICD-10 CM, FDT algorithm, MIDDM, and JASBO_ID-Net. This shows performance improvement of 7.03%, 4.60%, 2.30%, and 2.65%. Fig. 5 b) indicates recall-based comparison by changing training data. When training data is 50, recall is 0.854 for JASBO_ID-Net, wherein other models show recall of 0.781, 0.812, 0.837, as well as 0.828, with

improvement in the performance of 8.56%, 4.98%, 2.06%, and 3.13%. Figure 5 c) is an F-measure-based comparative assessment by changing training data. While the training percentage is 60, the F-measure is 0.849 for JASBO_ID-Net, wherein other models show an F-measure of 0.775, 0.786, 0.812, and 0.822. This shows performance enhancement of 8.77%, 7.42%, 4.43%, and 3.17%.

**Comparative Analysis by Varying k-value**

Fig.6 shows comparison of by changing k-value. Fig. 6 a) denotes accuracy based comparative assessment by varying k-value. While k-fold is 5, then accuracy is 0.810, 0.808, 0.823, 0.835, and 0.872 for DBN, ICD-10 CM, FDT algorithm, MIDDM , and JASBO_ID-Net, with performance improvement of 7.03%, 7.32%, 5.61%, and 4.20%. Figure 6 b) is recall based comparison by varying k-fold. While k-fold is 9, recall is 0.887 for JASBO_ID-Net, whereas other models show recall of 0.804, 0.849, 0.857, and 0.867, with performance improvement of 9.37%, 4.29%, 3.33%, and 2.22%. Fig.6 c) is F-measure based comparative analysis by changing k-value. While k-value is 7, then F-measure is 0.899 for JASBO_ID-Net, whereas other models show F-measure of 0.795, 0.819, 0.828, and 0.870. This shows enhancement in performance of 11.55%, 8.86%, 7.91%, as well as 3.17%.

**Comparative Discussion**

Table 1 resembles the comparative discussion of JASBO_ID-Net with various other methods like DBN, ICD-10 CM, FDT algorithm, and MIDDM. From table 1, it is found that JASBO_ID-Net exhibits high accuracy of 0.910, and this high value is because of Bi-LSTM that helped in extraction of character-based network features. Moreover, recall value is high of 0.887, and this is due to JASBO that helped in determining layer size. Also, F-measure is superior with value of 0.900, and this is because of ID-Net that was used in multi-classification of infectious disease from unstructured data. These superior values of performance metrics is attained while k-value is 9.
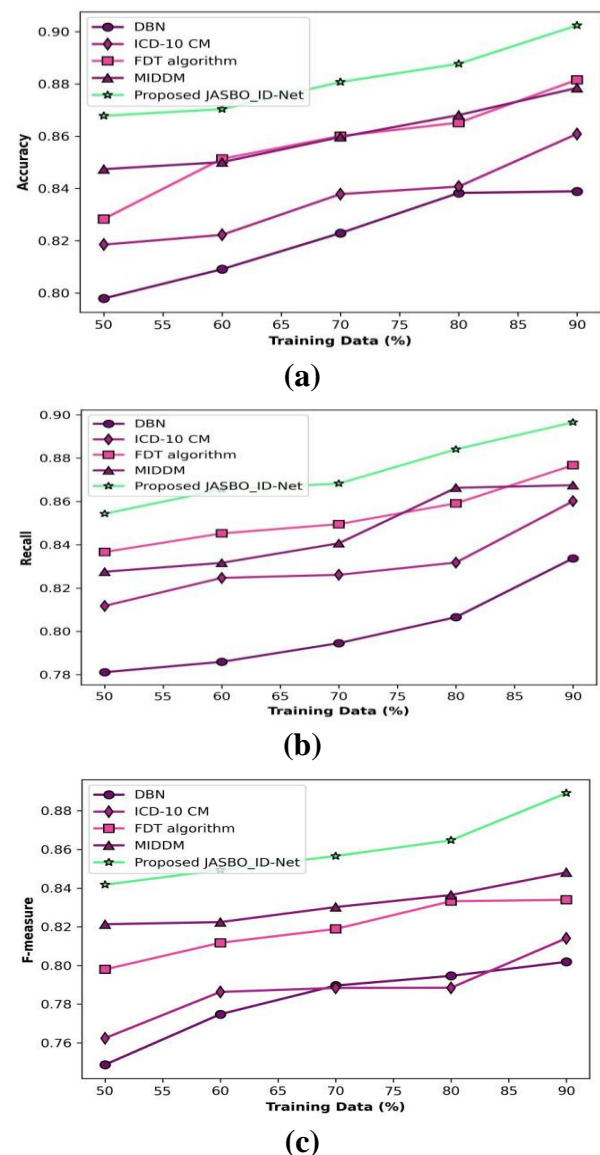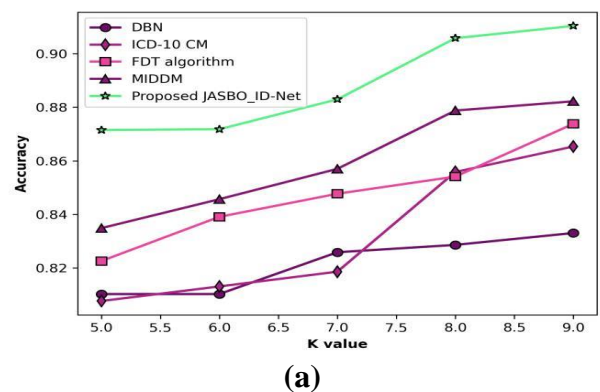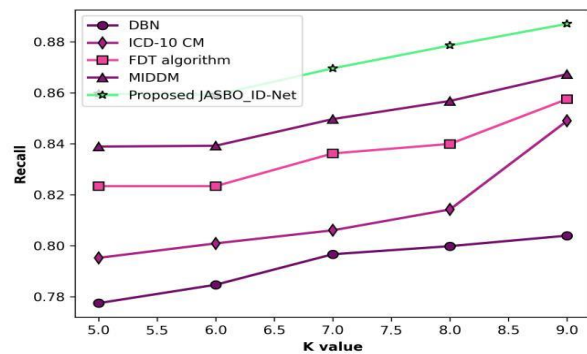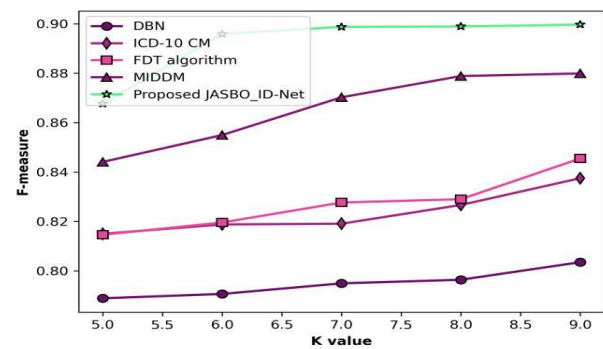


(a)



(b)



(c)

**Figure 5. Comparison of JASBO_ID-Net by changing training data, a) accuracy, b) recall, and c) F-measure**



(a)

**(b)**



**(c)**

**Figure 6. Comparison of JASBO_ID-Net by changing k-value, a) accuracy, b) recall, and c) F-measure**

**Table 1. Comparative discussion of JASBO_ID-Net**

| Classification | Metrics / Methods | DBN | ICD-10 CM | FDT algorithm | MIDDM | Proposed JASBO_ID-Net |
|---|---|---|---|---|---|---|
| **Training data=90%** | *Accuracy* | 83.9% | 86.1% | 88.2% | 87.8% | 90.2% |
| | *Recall* | 83.4% | 86.0% | 87.7% | 86.7% | 89.7% |
| | *F-measure* | 80.2% | 81.4% | 83.4% | 84.8% | 88.9% |
| **k-value=9** | *Accuracy* | 83.3% | 86.5% | 87.4% | 88.2% | **91.0%** |
| | *Recall* | 80.4% | 84.9% | 85.7% | 86.7% | **88.7%** |
| | *F-measure* | 80.3% | 83.7% | 84.5% | 88.0% | **90.0%** |

## Conclusion

Infectious diseases continue to be leading global cause of death, illness, disability, as well as socioeconomic unrest despite medical improvements. Early and accurate diagnosis and the right treatment option significantly influence each infection's outcome. Nowadays, there is a real need for a system that is able to multi-classify diseases by utilizing the various symptomatology present in the illnesses' unstructured texts. Medical unstructured script analysis is a crucial source for medical diagnosing of diseases. Accordingly, this paper contributes to medical informatics by presenting a proposed JASBO_ID-Net for multi-classifying infectious diseases. This work uses JASBO to determine layer size, which is formed by combining both Jaya algorithm and ASBO. Here, DL model used is ID-Net, which combines both CNN and Bi-LSTM. Initially, input is given in form of text to Token-ReLU, and is fed to Bi-LSTM layer. By Bi-LSTM model, character-based network features are extracted. Character-level features are then fed to the Attention layer, at which score function is calculated

by Kumar-Hassebrook similarity measure. The layer above the Bi-LSTM layer then predicts label of every word token. Here, JASBO combines the Average and Subtraction-Based Optimizer with Jaya algorithm. A number of performance metrics, including accuracy with a superior value of 91%, recall with a higher value of 88.7%, and F-measure with a superior value of 90%, are used to examine the best performance of the JASBO ID-Net. This research will be expanded in future for considering various issues associated with other forms of diseases like diabetes and heart diseases and to classify data using better algorithms for the benefit of society. We suggest for future studies to adopt models that may be more semantically explainable than models that depends on deep learning techniques. The researchers also can conducts test using external datasets that can be collected depending on IOT and wireless sensors to enhance models generalization and predictions. Also using non-English medical sources can be suggested for model development.

## Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of Middle Technical University.

## Authors' Contribution Statement

V. S. Performed conception and design, A. B. did the revision and proofreading, A. A. conducted the drafting of the MS.

## References

1. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. Front. Med. 2019 17;6:66. https://doi.org/10.3389/fmed.2019.00066.

2. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, Taylor RA. Neural Natural Language Processing for unstructured data in electronic health records: A review. Comput. Sci. Rev. 2022; 46:100511. https://doi.org/10.1016/j.cosrev.2022.100511.

3. Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, Kwak KS. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. inf. Fusion. 2020; 63:208-22. https://doi.org/10.1016/j.inffus.2020.06.008

4. Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, Cai T, Christiani DC. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA netw. Open. 2021; 4(7):e2114723-. https://doi.org/10.1001/jamanetworkopen.2021.14723

5. Arji G, Ahmadi H, Nilashi M, Rashid TA, Ahmed OH, Aljojo N, Zainol A. Fuzzy logic approach for infectious disease diagnosis: A methodical evaluation, literature and classification. BBE. 2019;39(4):937-55. https://doi.org/10.1016/j.bbe.2019.09.004

6. Moyo E, Mhango M, Moyo P, Dzinamarira T, Chitungo I, Murewanhema G. Emerging infectious disease outbreaks in Sub-Saharan Africa: Learning from the past and present to be better prepared for future outbreaks. Front. Public Health. 2023; 11:1049986. https://doi.org/10.3389/fpubh.2023.1049986

7. Bashir MF, Ma B, Shahzad L. A brief review of socio-economic and environmental impact of Covid-19. Air Qual. Atmos. Health. 2020; 13:1403-9. https://doi.org/10.1007/s11869-020-00894-8

8. Naz R, Gul A, Javed U, Urooj A, Amin S, Fatima Z. Etiology of acute viral respiratory infections common in Pakistan: A review. Rev. Med. Virol. 2019;29(2):e2024 https://doi.org/10.1002/rmv.2024

9. Wang M, Wei Z, Jia M, Chen L, Ji H. Deep learning model for multi-classification of infectious diseases from unstructured electronic medical records. BMC Med. Inform. Decis. Mak. 2022 ;22(1):1 https://doi.org/10.1186/s12911-022-01776-y

10. Luo X, Gandhi P, Zhang Z, Shao W, Han Z, Chandrasekaran V, Turzhitsky V, Bali V, Roberts AR, Metzger M, Baker J. Applying interpretable deep learning models to identify chronic cough patients using EHR data. Comput. Methods Programs Biomed. 2021; 210:106395. https://doi.org/10.1016/j.cmpb.2021.106395.

11. Vidhya K, Shanmugalakshmi R. Deep learning based big medical data analytic model for diabetes complication prediction. JAIHC. 2020; 11:5691-702. https://doi.org/10.1007/s12652-020-01930-2.

12. Wang SM, Chang YH, Kuo LC, Lai F, Chen YN, Yu FY, Chen CW, Li ZW, Chung Y. Using deep learning for automatic ICD-10 classification from free-text data. EJBI. 2020;16(1).

13. Zhao J, Yu L, Liu Z. Research based on multimodal deep feature fusion for the Auxiliary diagnosis model of Infectious Respiratory diseases. Sci. Program. 2021; 2021:1-6. https://doi.org/10.1155/2021/5576978.

14. Maheshwari V, Mahmood MR, Sravanthi S, Arivazhagan N, ParimalaGandhi A, Srihari K, Sagayaraj R, Udayakumar E, Natarajan Y, Bachanna P, Sundramurthy VP. Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control. J. Nanomater. 2021; 2021:1-8. https://doi.org/10.1155/2021/3383146.

15. Venkataraman GR, Pineda AL, Bear Don't Walk IV OJ, Zehnder AM, Ayyar S, Page RL, Bustamante CD, Rivas MA. FasTag: Automatic text classification of unstructured medical narratives. PLoS one. 2020; 15(6):e0234647. https://doi.org/10.1371/journal.pone.0234647.

16. Nagamine T, Gillette B, Pakhomov A, Kahoun J, Mayer H, Burghaus R, Lippert J, Saxena M. Multiscale

classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci. Rep.. 2020; 10(1):1-3. https://doi.org/10.1038/s41598-020-77286-6.

17. Ahmad A, Ullah A, Feng C, Khan M, Ashraf S, Adnan M, Nazir S, Khan HU. Towards an improved energy efficient and end-to-end secure protocol for IoT healthcare applications. Secur. Commun. Netw.. 2020; 2020:1-0. https://doi.org/10.1155/2020/8867792.

18. Ashraf S, Ahmed T, Aslam Z, Muhammad D, Yahya A, Shuaeeb M. Depuration based Efficient Coverage Mechanism for Wireless Sensor Network . J. Electr. Comput. Eng. Innovations. 2020; 8(2):145-60. https://doi.org/10.22061/jecei.2020.6874.344.

19. Ashraf S, Saleem S, Chohan AH, Aslam Z, Raza A. Challenging strategic trends in green supply chain management. Int. J. Res. Eng. Appl. Sci. JREAS. 2020; 5(2):71-4. https://doi.org/10.46565/jreas.2020.v05i02.006

20. Dehghani M, Hubálovský Š, Trojovský P. A new optimization algorithm based on average and subtraction of the best and worst members of the population for solving various optimization problems. PeerJ Comput. Sci.. 2022 ;8:e910. https://doi.org/10.7717/peerj-cs.910.

21. Venkata Rao R, Venkata Rao R. Jaya optimization algorithm and its variants. Jaya: An advanced optimization algorithm and its engineering applications. 2019:9-58. https://doi.org/10.1007/978-3-319-78922-4_2

22. MeDAL dataset , "https://www.kaggle.com/datasets/xhlulu/medal-emnlp", accessed on January 2023.

23. Sugave S, Jagdale B. Monarch-EWA: Monarch-earthworm-based secure routing protocol in IoT. Comput J. 2020; 63(6):817-31. https://doi.org/10.1093/comjnl/bxz135.

24. Hasan AM, Qasim AF, Jalab HA, Ibrahim RW. Breast Cancer MRI Classification Based on Fractional Entropy Image Enhancement and Deep Feature Extraction. Baghdad Sci. J. 2022; 20(1) :0221- 234. https://doi.org/10.21123/bsj.2022.6782

25. Li W, Qi F, Tang M, Yu Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. Neurocomputing. 2020; 387:63-77. https://doi.org/10.1016/j.neucom.2020.01.006

26. Kumar-Hassebrook similarity measure , https://drostlab.github.io/philentropy/reference/distance.html.

27. Wang SH, Muhammad K, Hong J, Sangaiah AK, Zhang YD. Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization. Neural Comput. Appl. 2020; 32:665-80. https://doi.org/10.1007/s00521-018-3924-0.

28. Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. J. Biomed. Inform.. 2020; 103:103381. https://doi.org/10.1016/j.jbi.2020.103381.

29. Wotaifi TA, Dhannoon BN. An Effective Hybrid Deep Neural Network for Arabic Fake News Detection. Baghdad Sci. J. 2023;20(4): https://doi.org/10.21123/bsj.2023.7427

30. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy. Nature. 2020; 585(7825):357-62. https://doi.org/10.1038/s41586-020-2649-2

31. Keras: Deep Learning for humans . [cited 2023 Aug 11]. https://keras.io/

32. Scikit-Learn: machine learning in Python — scikit-learn 1.3.0 documentation [Internet]. [cited 2023 Aug 11]. Available from: https://scikit-learn.org/stable/

33. Wen Z, Lu XH, Reddy S. MeDAL: medical abbreviation disambiguation dataset for natural language understanding pretraining. arXiv preprint arXiv:2012.13978. 2020. https://doi.org/10.48550/arXiv.2012.13978

# تشخيص الامراض الوبائية المتعددة من البيانات غير المبوبة باستخدام خوارزمية جايا لتحسين خوارزمية التعلم العميق

**فيان طلال صبيح ، أحمد بهاء الدين عبد الوهاب ، علي عبد المنعم ابراهيم الخراز**

قسم تقنيات المعلوماتية، الكلية التقنية الادارية/بغداد، الجامعة التقنية الوسطى، بغداد، العراق.

## الخلاصة

الأمراض الوبائية اصبحت مشكلة لا يمكن تجنبها في بيئتنا الحالية مع تميزها بنفس الاعراض الامر الذي يجعل من تشخيصها و الكشف المبكر عنها امراً صعباً. لذلك، اصبح من الضروري ايجاد تقنية تعتمد على اعراض المرضى المختلفة لتصنيف امراضهم. تعد الوثائق الطبية من المصادر المهمة التي مازالت تحتاج لطرق مبتكرة وموثوقة لتحليلها من اجل الوصول لتشخيص الكثير من الامراض، وعليه من المهم اضافه جهد في هذا المجال لأثراء معالجه النصوص الطبية للاستفادة منها في مجال المعلوماتية الصحية. ولهذا، تم افتراض خوارزمية JASBO اي خوارزمية تحسين معدل الفرق باستخدام JAYA التي تعتمد على التعلم العميق والتي تعمل على تصنيف الامراض المعدية الى فئاتها اعتمادا على البيانات النصية غير المبوبة في هذا البحث. ان شبكة تشخيص الامراض ID-NET التي افترضت تتكون من شبكة عصبية التفافية CNN من اجل تشخيص الكلمات الغريبة او الكلمات المفيدة في التشخيص مع شبكة الذاكرة الثنائية الاتجاه طويلة المدى BI-LSTM . حيث تم استخدام خوارزمية JASBO من اجل تحديد حجم الفلتر في شبكة التصنيف النهائية من اجل تحديد اهم اجزاء النص المعبرة عن المرض. يبدأ عمل الشبكة بدخول النص المطلوب تصنيفه الى مرحلة التقطيع الى كلمات، ليتم توجيه الكلمات لاحقاً شبكة تعلم عميق التفافية. اضافه لذلك يتم استخراج خصائص نمطيه او تراتبية احرف الكلمات باستخدام شبكة الذاكرة الثنائية لتتحول الى مصفوفه خصائص توجه الى طبقة تحسس لا يجاد تشابه تراتبية الحروف بين الكلمات بواسطة معادلة كومار-هانزبروك للتشابه . اعتماداً على ناتج شبكة JASBO يتم التنبؤ بفئة كل كلمة فيما اذا كانت تشير لأعراض مرض ما. الشبكة المفترضة لتشخيص الامراض أظهرت كفاءة بدقة 91 % ونسبة ارجاع 88% مع نسبة F-SCORE وصلت الى 90%.

**الكلمات المفتاحية:** خوارزمية تحسين معدل الفرق ، شبكة الذاكرة الثنائية ، الشبكة العصبية الالتفافية ، شبكة تشخيص اعراض الامراض ، خوارزمية جايا.