



Research Paper

Fast Decision Support for Infectious Diseases in Large Crowd Environments using NLP-Based Symptom-Driven Ensemble Learning Algorithms

Kavitha Kapala^{1*}, Vanjangi Jyotsna Devi², Laxmi Narasimha Krishna Sai Madupu³, Badepally Mallaiah⁴, Lakshmi Lalitha Sristi⁵, Prabagar Subbian⁶

1. Department of Computer Science and Engineering, GMR Institute of Technology (GMRIT), Deemed to be University, Rajam, Andhra Pradesh, India.

2. Department of Information Technology, Aditya Institute of Technology and Management, Srikakulam, Andhra Pradesh, India.

3. Department of Civil Engineering, R.V.R. & J.C. College of Engineering, Guntur, Andhra Pradesh, India.

4. Department of IT, CVR College of Engineering, Hyderabad, Telangana, India.

5. Department of CSE (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, India.

6. Department of Computer Science and Engineering (Data Science), Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India.

Citation Kapala K, Devi VJ, Madupu LNKS, Mallaiah B, Sristi LL, Subbian P. Fast Decision Support for Infectious Diseases in Large Crowd Environments using NLP-Based Symptom-Driven Ensemble Learning Algorithms. *International Journal of Medical Toxicology and Forensic Medicine*. 2026; 16:E52603.

<https://doi.org/10.22037/ijmtfm.v16.52603>

Article info:

Received: 26 June, 2026

Accepted: 13 July, 2026

Published: 19 September, 2026

Keywords:

Disease detection, Fast decision support, Deep learning, Natural language symptoms, Manta ray foraging optimization, Word2Vec

ABSTRACT

Background: The rapid growth of online medical data has increased the complexity of infectious disease detection. Natural language processing (NLP) is playing a vital part in transforming disease discovery by effectively analyzing clinical text data. This study develops a novel Fast Decision Support for Disease Detection in Large Crowd Environments with Heuristic Optimization Algorithm (FDSDD-LCEHOA) using Natural Language Symptoms.

Methods: For feature extraction, the Word2Vec model has been used to convert text into meaningful vector representations that capture contextual relationships among symptoms. For classification, an ensemble of deep learning (DL) techniques, namely bidirectional long short-term memory (BiLSTM), temporal convolutional networks (TCNs), and convolutional autoencoder (CAE), is employed to enhance predictive performance. Moreover, the hyperparameters of the three ensemble models are tuned using the manta ray foraging optimization (MRFO) algorithm to achieve optimal parameter settings.

Results: The proposed FDSDD-LCEHOA model achieved an accuracy of 98.37%, demonstrating its effectiveness in classifying diseases from natural-language symptoms.

Conclusion: The experimental results show that the proposed model outperforms existing models, enabling real-time disease monitoring and early outbreak detection in large-crowd environments.

* Corresponding Author:

Kavitha Kapala, PhD

Associate Professor, Department of Computer Science and Engineering, GMR Institute of Technology (GMRIT) - Deemed to be University, Rajam, Andhra Pradesh - 532127, India.

E-mail: kavitha.akkil1514@gmail.com



Copyright © 2026 The Author(s).

This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC: <https://creativecommons.org/licenses/by-nc/4.0/legalcode.en>), which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Introduction

Heavily crowded zones are likewise conducive to the emergence and transmission of certain respiratory outbreaks [1]. Repeated interactions among individuals, whose bodily contact escalates with population density in any given area, led to the spread of infectious diseases to a large group in a shorter period. Healthcare is one of the most pressing issues in human communities. Nevertheless, the medical field is highly varied, widely dispersed, and fragmented [2]. Early disease detection is crucial and life-preserving, and it is performed using both non-invasive and invasive approaches. However, Artificial intelligence (AI) offers innovative tools for identifying signs of disorder and novel treatment techniques for monitoring the disorder's progression. Deep learning (DL), combined with disease detection, was applied in a monitoring setup to support human examination.

The following section provides a detailed study of existing methods in the field of infectious disease detection using various ML, DL, and NLP. Ravi et al. [3] presented a new system for the early detection of pancreatic cancer. Using cutting-edge DL models, particularly a combination of SVM and MLP. Amannah et al. [4] presented a prototype diagnostic method that incorporates XAI and ML to enhance diagnostic output, understandability, and utility in resource-limited environments. Qiang et al. [5] proposed a new 3D patch-level AD prediction model, 3DSC-TF, comprising a 3DSCNN and a transformer model to achieve accurate AD detection. Rehman et al. [6] focused on dysphonia, a modification in the vocal sound. Voice syndromes become prevalent despite being frequently unnoticed. Mobile healthcare applications could provide rapid, effective support for detecting voice syndromes. The author [7] aimed to classify ASD and typically established TD class data through ML classifications. Talukder et al. [8] introduced a hybrid explainable AI-aided ML (HXAI-ML) system for addressing those restrictions. Inthiyaz et al. [9] provided an automated image-based model for identifying and classifying skin conditions using ML. Kundu and Anguraj [10] proposed an OML-AMPDC system using blood images. This suggested system mainly performs preprocessing in two phases. Furthermore, the feature extraction step is applied using Local Derivative Radial Patterns (LDRP)

Consequently, this study develops an FDSDD-

LCEHOA using Natural Language Symptoms. The primary goal is to design an intelligent, fast decision-support system for the early identification of infectious disease outbreaks in large-crowd environments using NLP and optimized ensemble deep-learning approaches. The important contributions of this study are summarized as given below:

- The Word2Vec-based feature extraction process has been applied to capture contextual relationships among symptom descriptions effectively.
- An ensemble of DL approaches such as bi-directional long short-term memory (Bi-LSTM), temporal convolutional networks (TCNs), and convolutional autoencoder (CAE) models is employed for better disease classification.
- The hyperparameter tuning can be accomplished through the manta ray foraging optimization (MRFO) algorithm to improve the model accuracy and efficiency.

Materials and Methods

Cell Line and Culture Conditions

In this paper, we developed a novel FDSDD-LCEHOA method to support large-scale disease detection and early identification of infectious diseases. The model consists of several stages, including pre-processing, feature extraction, ensemble classification, and parameter selection, to enhance performance and ensure reliable decision-making. Figure 1 illustrates the workflow of the proposed FDSDD-LCEHOA model.

Feature Extraction using Word2Vec

To perform the feature extraction phase, the Word2Vec framework has been employed to convert text into meaningful vectors that capture contextual relationships among symptoms. The traditional word embedding model utilized unsupervised learning methods such as LSA, skip-gram, and related models. Moreover, it was chosen because it efficiently captures semantic relationships among symptom-related terms while maintaining low computational complexity. Compared with transformer-based embedding models such as BioBERT and ClinicalBERT, Word2Vec requires fewer computational resources and shorter training time, making it suitable for the proposed disease classification framework. Although contextual language models can provide richer contextual representations, they typically require large-scale, domain-specific pretraining and incur higher computational costs. Therefore, Word2Vec was adopted in this study to achieve an effective balance

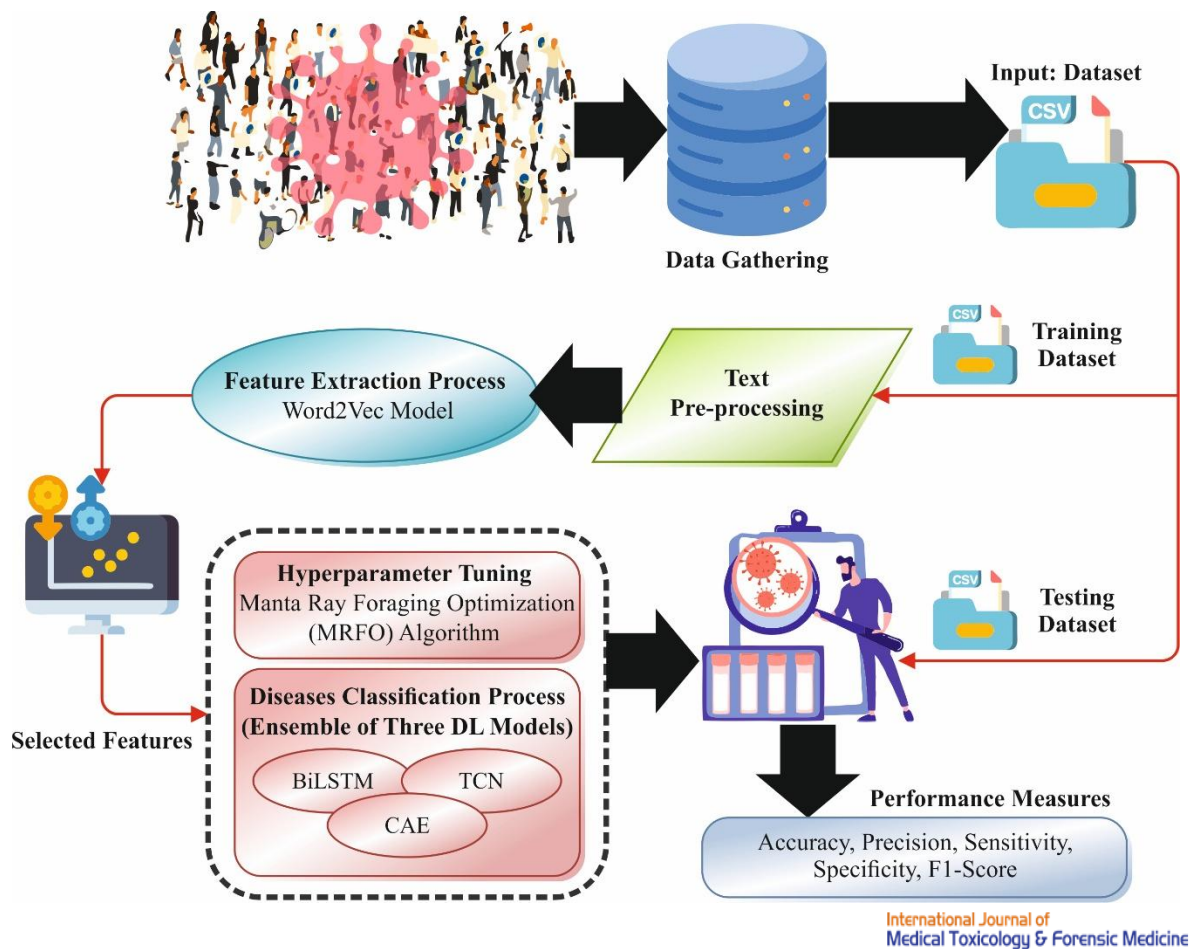


Figure 1. Complete flow of FDSDD-LCEHOA model.

between semantic representation and computational efficiency.

Furthermore, Compared with BioBERT and ClinicalBERT, Word2Vec requires lower computational resources, less memory, and shorter training time while still providing effective semantic representations of disease-related text. Its computational efficiency and simple implementation make it a suitable choice for the proposed disease classification framework. Advanced word representations utilizing the so-called word2vec, utilizing an NN language method trained in dual phases.

Ensemble Learning Models

The ensemble framework integrates the prediction scores from the BiLSTM, TCN, and CAE models via a decision-level aggregation strategy. Each model independently processes the optimized Word2Vec feature vectors and produces a probability score for each disease class. Since all three models are considered equally important in the proposed framework, an equal-weight average is used to obtain the final prediction score. The aggregated score is computed as:

$$S_{\text{final}} = \frac{1}{3} (S_{\text{BiLSTM}} + S_{\text{TCN}} + S_{\text{CAE}}) \quad (1)$$

where S_{final} denotes the final aggregated prediction score, while S_{BiLSTM} , S_{TCN} , and S_{CAE} represent the prediction scores generated by the BiLSTM, TCN, and CAE models, respectively. The factor $\frac{1}{3}$ assigns equal importance to each model, ensuring that all three classifiers contribute uniformly to the ensemble decision.

BiLSTM Classifier

One kind of RNN that can identify longer-term dependencies in the input sequence is a Bi-LSTM. This implies that, unlike conventional methods, the Bi-LSTM learns the preceding and succeeding contexts from the given data. To introduce how BiLSTM systems process sequential data and improve prediction, it upgrades either forward or Backward LSTM Cells

TCN Classifier

While CNNs exhibit outstanding feature-extraction capabilities, they struggle to process long time-series

data due to vanishing gradients. In this context, TCNs are presented as an advanced framework based on CNNs to enhance performance in addressing longer time-series data. Dilated convolution increases the spacing between kernel elements within the filter, thereby expanding the receptive field of the model without a significant increase in parameters. The final output before the activation function of the dilated-causal convolution is stated as:

$$z_{t,n}^{cl} = \sum_{k=0}^{N_t-1} x(t - k \cdot d)w_{t,n}(k) + b_{t,n} \quad (8)$$

The dilated-causal convolution expands the receptive field and mitigates the vanishing gradient problem by establishing direct input-output connections, as presented in Eq. (8).

$$y_t^r = y_t^{r-1} + \sigma_t(z_t^{cl,r}) \quad (9)$$

The training procedure of TCN additionally comprises dual sections: backward and forward propagation.

$$y_{t,n}^{cl} = \sigma_t(z_{t,n}^{cl}) = \text{ReLU}(z_{t,n}^{cl}) \quad (10)$$

$$y_t^r = \sigma_t(z_t^{cl,r}) + y_p^{r-1} \quad (11)$$

$$y_t^{ml} = \sigma_t(z_t^{ml}) = \sigma_t(w_t^{ml}y_t^{ml-1} + b_t^{ml}) \quad (12)$$

Whereas $w_t^{ml} = \{w_{t,1}^{ml}, w_{t,2}^{ml}, \dots, w_{t,N_{ml}-1}^{ml}, w_{t,N_{ml}}^{ml}\}$ denote weighting vectors of the ml -th fully connected (FC) layer of TCN and b_t^{ml} means bias of the ml -th FC layer.

CAE Classifier

The CAE approach is employed in the classification technique. Diverse CAE employs Conv layers to process composite data, primarily images. This architecture allows the CAE to effectively investigate substantial outcomes using the spatial region of aspects. The AE is ANN, integrating a 3-layer progression connected with sequence, precisely output, input, and HLs. The two phases can be given:

$$h(a) = f(W_1a + b_1) \quad (13)$$

$$A = g(W_2h(a) + b_2) \quad (14)$$

Here, $h(a)$ is the encoder output, A is the decoder output, f and g are the encoder and decoder functions, and W_1 , W_2 , b_1 , and b_2 denote the weights and bias vectors (Figure 2).

Fine-Tuning Parameter using MRFO

The hyperparameter tuning for the 3 ensemble methods is performed using the MRFO method to achieve optimal parameter tuning. The MRFO algorithm is stimulated by 3 hunting behaviors of manta rays [11]. The MRFO model employs the fitness function (FF) based on the classification error rate, as given in Eq. (24).

$$\begin{aligned} \text{fitness}(x_i) &= \text{ClassifierErrorRate}(x_i) \\ &= \frac{\text{no. of misclassified instances}}{\text{Total number of instances}} * 100 \end{aligned} \quad (24)$$

The convergence behavior of MRFO was monitored throughout the optimization process. The fitness value gradually improved over successive iterations and stabilized after several optimization cycles. The computational complexity of the MRFO optimization process is primarily dependent on the population size (N), the maximum number of iterations (T), and the cost of evaluating the objective function (C_f). Accordingly, the overall time complexity can be expressed as:

$$O(N \times T \times C_f) \quad (25)$$

Although the optimization stage introduces additional computational overhead during training, it is performed only once and does not affect the inference efficiency of the proposed disease classification framework.

Results

The empirical assessment of the FDSDD-LCEHOA model has been examined in the Symptom2Disease dataset [12] and the Symptom to Disease Descriptions dataset [13]. This dataset contains 350 total symptoms (Table 1).

Table 2 presents a comparison of the FDSDD-LCEHOA technique with recent algorithms across various measures on the Symptom2Disease dataset [14, 15]. The result implied that the CNN, LSTM, BERT-FT, MediGPT Improvement, GBM, MLP, RSVM, and XGBOOST approaches perform minimal solutions compared to the proposed method, which achieves maximal value with $accu_y$ of 98.37%, $prec_n$ of 93.79%, $F1_{score}$ of 93.71%, $sens_y$ of 93.91%, and $spec_y$ of 93.91%, correspondingly (Table 2).

Figure 3 presents a comparative study of the FDSDD-LCEHOA technique against recent algorithms

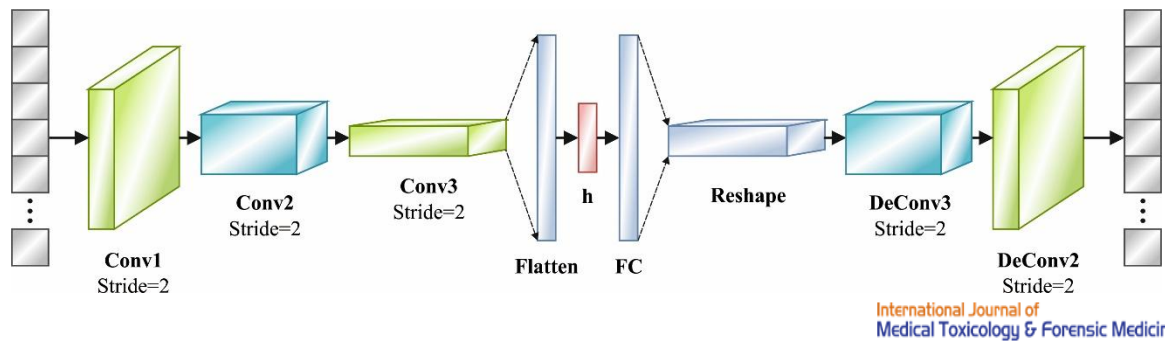


Figure 2. Framework of CAE.

Table 1. Parameter setting.

Method	Parameter	Value
Text Preprocessing	Maximum Sequence Length	100
Word2Vec	Architecture	Skip-Gram
Word2Vec	Embedding Dimension	300
Word2Vec	Context Window	5
BiLSTM	Hidden Units	128
BiLSTM	Number of Layers	2
TCN	Filters	128
TCN	Kernel Size	3
TCN	Dilation Rates	1, 2, 4, 8
CAE	Latent Dimension	128
Ensemble	Fusion Strategy	Soft Voting
MRFO	Population Size	30
MRFO	Maximum Iterations	100
Optimizer	Algorithm	Adam
Learning Rate	Initial LR	0.001
Batch Size	Batch Size	32
Training	Epochs	50

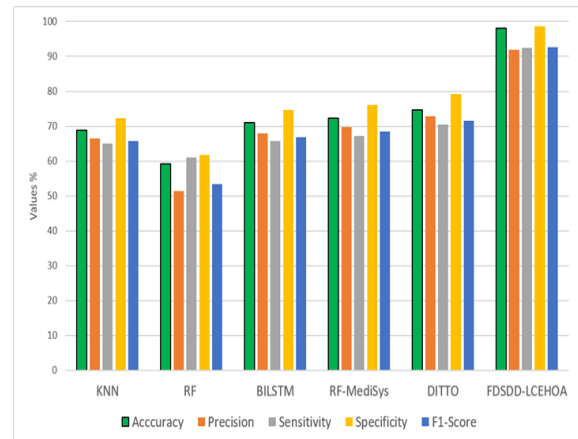
International Journal of
Medical Toxicology & Forensic Medicine

Table 2. Comparison evaluation of the FDSDD-LCEHOA method with recent algorithms under the Symptom2Disease dataset.

Algorithm	$Accu_y$	$Prec_n$	$Sens_y$	$Spec_y$	$F1_{score}$
CNN	92.53	90.05	90.01	89.28	90.40
LSTM	89.77	89.36	91.67	98.62	90.60
BERT-FT	95.87	90.78	91.99	95.22	92.39
MediGPT Improvement	97.99	91.92	89.75	93.28	90.77
GBM	95.30	91.57	92.09	89.54	90.39
MLP	98.04	91.17	90.57	98.60	93.44
RSVM	90.15	92.63	92.25	95.50	92.16
XGBOOST	94.81	92.54	92.53	98.62	90.99
FDSDD-LCEHOA	98.37	93.79	93.91	99.05	93.71

International Journal of
Medical Toxicology & Forensic Medicine

across various measures on the Symptom to Disease Descriptions dataset [16, 17]. The result implied that the KNN, RF, BiLSTM, RF-MediSys, and DITTO techniques obtain the least performance, compared with the presented approach, which achieves a better



International Journal of
Medical Toxicology & Forensic Medicine

Figure 3. Comparative analysis of the proposed FDSDD-LCEHOA methodology under the Symptom to Disease Descriptions data set with existing methods.

value $accu_y$ of 98.05%, $prec_n$ of 91.83%, $sens_y$ of 92.41%, $spec_y$ of 98.63%, and $F1_{score}$ of 92.54%, correspondingly.

Discussion

Table 3 shows the computational efficiency of the FDSDD-LCEHOA model with existing models. Based on inference time, the FDSDD-LCEHOA method provides lesser value of 5.62ms, Flops of 0.64G and GPU of 0.68GB, with other models namely, the CNN, LSTM, BERT-FT, MediGPT Improvement, GBM, MLP, RSVM, and XGBOOST achieves a greater time of 12.84ms, 15.72ms, 28.65ms, 31.43ms, 8.47ms, 7.95ms, 9.18ms, and 6.84ms, correspondingly.

Table 4 presents the K-fold cross-validation results for the proposed FDSDD-LCEHOA model. In K 1, the model achieved an Accuracy of 96.84%, Precision of 90.28%, Recall of 90.91%, and F1 Score of 90.59%. Finally, in K=10, the model achieved an Accuracy of 97.11%, Precision of 91.28%, Recall of 91.86%, and F1 Score of 91.57%, confirming stable and consistent performance in the final fold of the cross-validation process.

The proposed model achieves superior performance

Table 3. Computational efficiency of the FDSDD-LCEHOA method with existing techniques.

Methods	Inference Time (ms)	FLOPs (G)	GPU Memory (GB)
CNN	12.84	3.46	2.18
LSTM	15.72	4.28	2.74
BERT-FT	28.65	22.84	7.96
MediGPT Improvement	31.43	25.63	8.84
GBM	8.47	1.12	0.98
MLP	7.95	0.86	0.81
RSVM	9.18	1.34	1.05
XGBOOST	6.84	0.79	0.73
FDSDD-LCEHOA	5.62	0.64	0.68

International Journal of
Medical Toxicology & Forensic Medicine**Table 4.** K-fold cross-validation of the proposed FDSDD-LCEHOA approach.

K-Fold	Accuracy	Precision	Recall	F1-Score
K-1	96.84	90.28	90.91	90.59
K-2	97.23	91.46	91.02	91.24
K-3	96.51	89.82	89.35	89.58
K-4	97.48	92.08	91.73	91.9
K-5	96.92	90.94	91.28	91.11
K-6	97.36	91.67	92.11	91.89
K-7	97.82	92.71	93.04	92.74
K-8	96.73	90.53	90.07	90.3
K-9	97.61	92.34	92.45	92.39
K-10	97.11	91.28	91.86	91.57

International Journal of
Medical Toxicology & Forensic Medicine

in analyzing unstructured medical text in crowd environments, delivering strong, consistent results in infectious disease detection and early outbreak identification.

Conclusion

In this manuscript, we have developed a new FDSDD-LCEHOA approach using Natural Language Symptoms. The FDSDD-LCEHOA methodology supports scalable disease detection and public health monitoring in high-density environments. In the feature extraction process, the Word2Vec framework has been employed to convert text into meaningful vectors that capture contextual relationships among symptoms. Also, an ensemble model is used for disease classification. Eventually, hyperparameter tuning for the 3 ensemble systems is performed using the MRFO model to optimize the parameters. To determine the effective classification result of the FDSDD-LCEHOA system, a broad range of simulations are conducted on the benchmark dataset. The wide-ranging comparative

outcomes demonstrated the benefits of the FDSDD-LCEHOA model using recent techniques. Future research will focus on integrating contextual biomedical language models, developing uncertainty-aware prediction mechanisms, evaluating the framework across diverse clinical datasets, and conducting large-scale deployment studies to enhance robustness, scalability, and practical applicability in healthcare environments.

Ethical Approval

Not Applicable.

Acknowledgment

Noe.

Funding

No external funding was received for this research.

Conflicts of Interest

The authors report there are no competing interests to declare.

References

- [1] Sreenu GSDMA, Durai S. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J Big Data*. 2019;6(1):1-27. [DOI: 10.1186/s40537-019-0212-5]
- [2] Tripathi G, Singh K, Vishwakarma DK. Convolutional neural networks for crowd behaviour analysis: a survey. *Vis Comput*. 2019;35(5):753-76. [DOI: 10.1007/s00371-018-1499-5]
- [3] Ravi KC, Dixit RR, Singh S, Gopatoti A, Yadav AS. AI-powered pancreas navigator: Delving into the depths of early pancreatic cancer diagnosis using advanced deep learning techniques. In: 2023 9th International Conference on Smart Structures and Systems (ICSSS). IEEE; 2023. p. 1-6. [DOI: 10.1109/ICSSS58416.2023.10494488]
- [4] Amannah C, Attai KF, Uzoka FM. A Data-Driven Intelligent Methodology for Developing Explainable Diagnostic Model for Febrile Diseases. *Algorithms*. 2025;18(4):190. [DOI: 10.3390/a18040190]
- [5] Qiang YR, Zhou QY, Li JN, Xie MY, Cui X,

- Zhang SW. Classification of Alzheimer's disease by jointing 3D depthwise separable convolutional neural network and transformer. *Expert Syst Appl.* 2025;274:127720. [DOI: 10.1016/j.eswa.2025.127720]
- [6] Rehman MU, Shafique A, Azhar QUA, Jamal SS, Gheraibia Y, Usman A. Voice disorder detection using machine learning algorithms: An application in speech and language pathology. *Eng Appl Artif Intell.* 2024;133(Pt A):108047. [DOI: 10.1016/j.engappai.2024.108047]
- [7] Shrivastava T, Singh V, Agrawal A. Autism spectrum disorder detection with kNN imputer and machine learning classifiers via questionnaire mode of screening. *Health Inf Sci Syst.* 2024;12(1):18. [DOI: 10.1007/s13755-024-00277-8]
- [8] Talukder MA, Talaat AS, Kazi M. HXAI-ML: a hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection. *Results Eng.* 2025;25:104370. [DOI: 10.1016/j.rineng.2025.104370]
- [9] Inthiyaz S, Altahan BR, Ahammad SH, Rajesh V, Kalangi RR, Smirani LK, et al. Skin disease detection using deep learning. *Adv Eng Softw.* 2023;175:103361. [DOI: 10.1016/j.advengsoft.2022.103361]
- [10] Kundu TK, Anguraj DK. Optimal Machine Learning Based Automated Malaria Parasite Detection and Classification Model Using Blood Smear Images. *Trait Signal.* 2023;40(1):319-27. [DOI: 10.18280/ts.400108]
- [11] Abdul Razak AA, Nasir ANK, Abdul Ghani NM, Ahmad MR, Jui MJJ. Efficient Constrained Real-World Problem-Solving Using Gradient-Based Mutation Manta Ray Foraging Optimization (GM-MRFO). *IEEE Access.* 2026;14. [DOI: 10.37934/araset.61.1.138157]
- [12] Barman N. Symptom2Disease [Internet]. Kaggle; 2023 [cited 2026 Sep 16]. [Link]
- [13] Roy BA. Symptom to disease descriptions 820 cleaned [Internet]. Kaggle; 2023 [cited 2026 Sep 16]. [Link]
- [14] Rony MAT, Islam MS, Sultan T, Alshathri S, El-Shafai W. MediGPT: Exploring Potentials of Conventional and Large Language Models on Medical Data. *IEEE Access.* 2024;12:115206-20. [DOI: 10.1109/ACCESS.2024.3428918]
- [15] Hoang TH, Zallek C, Do MN. Smartphone-Based Digitized Neurological Examination Toolbox for Multi-test Neurological Abnormality Detection and Documentation. *IEEE J Biomed Health Inform.* 2024;28(10):5888-99. [DOI: 10.1109/JBHI.2024.3439492]
- [16] Ghosh S. Symptom-based disease prediction using deep learning with uncertainty estimation. *J Decis Anal Intell Comput.* 2026;6(1):32-50. [DOI: 10.31181/jdaic612026108]
- [17] Xu T, Deng X, Meng X, Yang H, Wu Y. Clinical NLP with attention-based deep learning for multi-disease prediction. In: 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC). *IEEE;* 2025. p. 382-6. [DOI: 10.48550/arXiv.2507.01437]