IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

Predict the Disease Based on Symptoms Using RNN Techniques

Abinaya T, Kaneshiya S, Jeya Ramya V

U.G. Student, Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Chennai, India

U.G. Student, Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Chennai, India

Professor, Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Chennai, India

ABSTRACT: Predicting diseases based on symptoms using Natural Language Processing (NLP) techniques is a burgeoning field with significant potential for improving healthcare diagnostics. NLP enables the analysis and interpretation of unstructured text data, such as patient symptoms recorded in electronic health records or online health forums. This abstract explores the application of NLP techniques in disease prediction based on symptoms, highlighting the benefits, challenges, and future directions of this approach. The abstract begins by introducing the importance of accurate disease prediction and the role of NLP in processing textual symptom data. NLP techniques enable the extraction of relevant information from symptom descriptions, such as the presence, severity, and temporal aspects of symptoms. These techniques also facilitate the identification of symptom co-occurrences and relationships, which can aid in disease prediction.

KEYWORDS: Disease prediction, Symptoms, Natural Language Processing (NLP), Text analysis, Named Entity Recognition (NER), Semantic Role Labeling (SRL), Sentiment analysis, Medical gerontologist, Deep learning, Transfer learning, Multi-modal data.

I. INTRODUCTION

The advent of Natural Language Processing (NLP) techniques has ushered in a trans-formative era in healthcare, offering innovative approaches to disease prediction and diagnosis. In recent years, there has been a growing interest in harnessing the power of NLP to predict diseases based on symptomatic information. This paradigm shift represents a convergence of computational linguistics, medical expertise, and machine learning, aiming to enhance the accuracy and efficiency of disease identification.Traditionally, disease prediction has relied on structured data, such as lab results and medical histories. However, the incorporation of NLP allows us to extract valuable insights from unstructured textual data, such as electronic health records, clinical notes, and patient-reported symptoms. This unstructured data often contains nuanced information, reflecting the subtleties of patient experiences that might elude structured approaches.The complexity of human language, with its variability, context-dependency, and ambiguity, presents both a challenge and an opportunity. NLP techniques enable the development of models capable of understanding and interpreting the intricate relationships between symptoms and underlying medical conditions. By leveraging advanced machine learning algorithms, these models can discern patterns, associations, and dependencies within textual data, empowering healthcare professionals with timely and accurate disease predictions.

II. EXISTING SYSTEM

This systematic review surveyed automatic emotion recognition systems applied in real clinical contexts, focusing on populations with pathologies. The review included 52 scientific papers meeting the inclusion criteria. Clinical applications primarily targeted neuro developmental, neurological, and psychiatric disorders, aiming to diagnose, monitor, or treat emotional symptoms. Observational study designs were common for monitoring and diagnosis, while interventional approaches were used for treatment. Video and audio signals were the most adopted, and supervised shallow learning was the prevalent approach for emotion recognition algorithms. Clinical limitations included small

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

sample sizes, absence of control groups, and lack of real-life conditions testing. Technically, heterogeneity in performance metrics, datasets, and algorithms challenged result comparability, robustness, reliability, and reproductibility. Suggested guidelines were provided to address these challenges and guide future research.

III. LITERATURE SURVEY

Title: NLP based Segmentation Protocol for Predicting Diseases and Finding Doctors
Author: Aswathy K P1, Rathi R2, Shyam Shankar E P3
Year: 2019
-In these decades the Web-browsing have become a habitual or imperative thing for the mankind. The people have been using many browsers, online/offline Apps and Software for their daily needs including their healthcare queries. The proper clarification of such healthcare queries and accurate disease-predictions will be very serviceable for the society. In light of this we developed a medical query system which has the ability not only to predict the disease by analyzing the user's symptoms but also to suggest doctors in user's locality if entered the locale while registering. In this paper we proposed a NLP based Unstructured Data Processing Method (NUDPM) which enables the facile segmentation and extraction of medical terms/symptoms from the user queries. The analysis of this NUDPM computed data with a predefined 'Disease data base' will predict the disease accurately. As well as the execution of Doctor-Type selection methods will choose appropriate doctors near to user's locality. Finally the system exhibits the details of the disease and nearby doctors to user.

Title: Multiple Disease Prognostication Based On Symptoms Using Machine Learning Techniques
Author: Kajal Patil , Sakshee Pawar , Pramita Sandhyan and Jyoti Kundale
Year: 2022
Disease Prediction system that uses Machine Learning forecasts the ailments on the basis of the data pertaining to the symptoms entered by the user and provides trustworthy findings based on that data. If the patient isn't in any danger and the user merely wants to know what kind of ailment he or she has had. It is a system that gives the user suggestions and methods on how to keep their health system in good shape, as well as a way to find out if they have a sickness utilizing this forecast. Due to a diversity of diseases and a lower doctor-patient ratio, the use of particular disease prediction technologies as well as concerns about health has risen. We are focusing on offering customers with an instant and accurate disease prognosis based on the symptoms they enter, as well as the severity of the condition projected. It will provide the best algorithm and doctor consultation. Different machine learning algorithms are employed to forecast illnesses, ensuring speedy and reliable predictions

IV. PROPOSED SYSTEM

The proposed system aims to leverage NLP techniques for disease prediction based on symptoms, incorporating advancements in the field to enhance accuracy and usability. The system consists of several components and methodologies to achieve its objective. The key components of the proposed system are as follows. Data Collection: The system will gather a comprehensive data set comprising symptom descriptions along with their corresponding disease labels. This data set can be obtained from electronic health records, patient forums, medical literature, or crowd sourced platforms. The data set will serve as the foundation for training and evaluating the disease prediction model. Pre-processing and Text Analysis: The system will employ NLP techniques for pre-processing and text analysis. This includes tasks such as tokenization, part-of-speech tagging, and syntactic parsing to extract relevant information from symptom descriptions. Named Entity Recognition (NER) will be applied to identify medical entities (symptoms, body parts, medical conditions) mentioned in the text. Semantic Role Labeling (SRL) will help understand the relationships between symptoms and modifiers, aiding in accurate symptom interpretation.

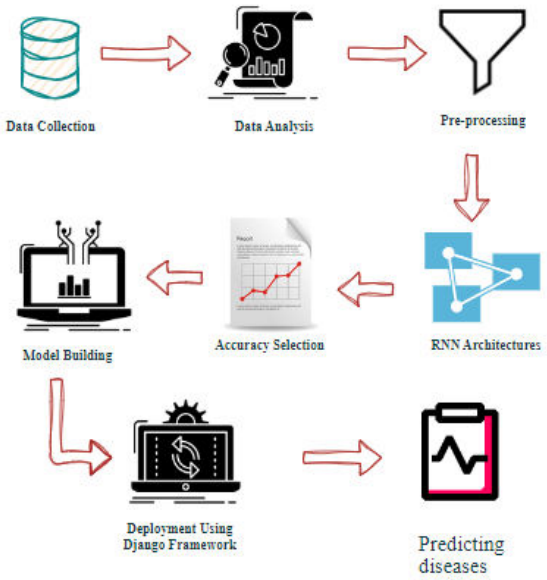
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

SYSTEM ARCHITECTURE



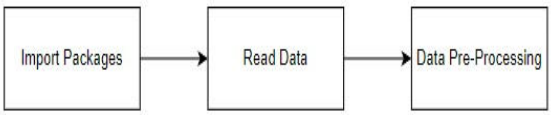
V. LIST OF MODULES

1. Data Preprocessing
2. Data Analysis of Visualization
3. GRU
4. LSTM
5. Deployment

MODULE DESCRIPTION:

Data Preprocessing:

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the data set. If the data volume is large enough to be representative of the population, you may not need the validation techniques.To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training data set while tuning model hyper parameters.The evaluation becomes more biased as skill on the validation data set is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation.Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.



International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

Data visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a data set and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance

Algorithm implementation:

The below 2 different algorithms are compared:
1.GRU Architecture
2.Long-Short term memory networks

GATED RECURRENT UNIT ARCHITECTURE:

Implementing the Gated Recurrent Unit (GRU) architecture involves utilizing a type of recurrent neural network (RNN) that incorporates gating mechanisms to manage information flow. By combining reset and update gates, GRU efficiently captures long-range dependencies in sequential data, making it suitable for tasks like Predicting diseases where contextual understanding is crucial for accurate classification.

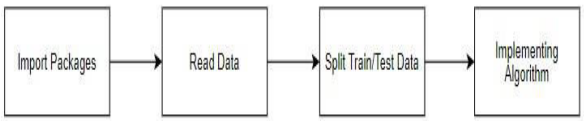
Basic Structure:
An RNN consists of a series of interconnected layers. At each time step t, it takes an input vector (or sequence) and produces an output vector (or sequence).
The key feature of an RNN is its hidden state, denoted as "h." This hidden state is a representation of the network's memory, and it is updated at each time step.

Input and Output:
At each time step t, the RNN takes an input vector or element x(t). This input can be a single element of a sequence, a word in a sentence, a pixel in an image, etc.
The RNN produces an output vector or element y(t) at each time step. The output can be used for various tasks, such as predicting the next element in a sequence or classifying the sequence as a whole.

Hidden State:
The hidden state h(t) is a vector that captures information from previous time steps. It serves as the memory of the network.
The hidden state is computed at each time step using the current input x(t) and the previous hidden state h(t-1).

Output Computation:
The output at each time step can be computed based on the current hidden state or a combination of the hidden state and the input at that time step.

Back propagation Through Time (BPTT):
Training an RNN involves using a variant of back propagation called Back propagation through Time.



International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

LSTM ARCHITECTURE:

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem and capture long-term dependencies in sequential data. It consists of memory cells, input gates, forget gates, and output gates. The memory cells store information over long sequences, while the gates regulate the flow of information, allowing LSTM to effectively learn and remember patterns in time-series data. In the context of Predicting diseases architecture helps the model understand the context of user comments and detect spam based on intricate patterns and dependencies within the text.

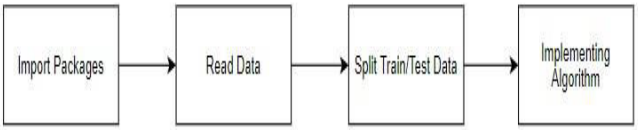
Basic Structure:
An LSTM network is composed of LSTM cells arranged in a sequence. Each LSTM cell has an internal structure that enables it to store and retrieve information over long sequences.
Like standard RNN, LSTM networks take input vectors or elements sequentially and produce output vectors or elements at each time step.
The key innovation in LSTM cells is their ability to maintain a cell state, which can capture long-term dependencies in the data.

Components of an LSTM Cell:
An LSTM cell consists of three main gates and a cell state:
Forget Gate: Decides what information from the cell state should be thrown away or kept.
Input Gate: Determines what new information should be added to the cell state.
Output Gate: Controls what information from the cell state should be used to generate the output.
Cell State: The cell state runs throughout the entire sequence and can carry information over long distances.

Information Flow:
The forget gate (f(t)) controls what information from the previous cell state (C(t-1)) should be retained.
The input gate (i(t)) determines what new information from the candidate cell state (ĉ(t)) should be added to the cell state.
The cell state (C(t)) is updated based on the forget gate, input gate, and candidate cell state.
The output gate (o(t)) controls what information from the cell state should be used to produce the hidden state (h(t)).

Back propagation Through Time (BPTT):
LSTM networks are trained using Back-propagation Through Time, similar to standard RNN. BPTT computes gradients for the network's parameters to minimize a loss function.

Advantages of LSTM:
LSTM can capture long-range dependencies in sequences.
They mitigate the vanishing gradient problem, allowing for more effective training on long sequences.
They are suitable for a wide range of sequence-based tasks and have been extended into more advanced variants like Gated Recurrent Units (GRUs)**.**



Deployment:
Django (Web Framework):
Django is a micro web framework written in Python.It is classified as a micro-framework because it does not require particular tools or libraries.It has no database abstraction layer, form validation, or any other components where preexisting third-party libraries provide common functions.However, Django supports extensions that can add application features as if they were implemented in Django itself.
Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

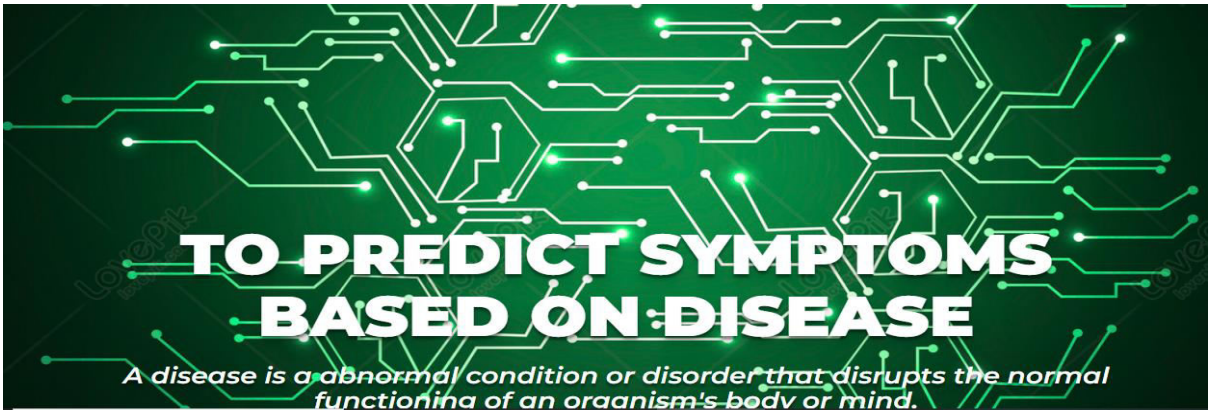
International Journal of Innovative Research in Computer and Communication Engineering

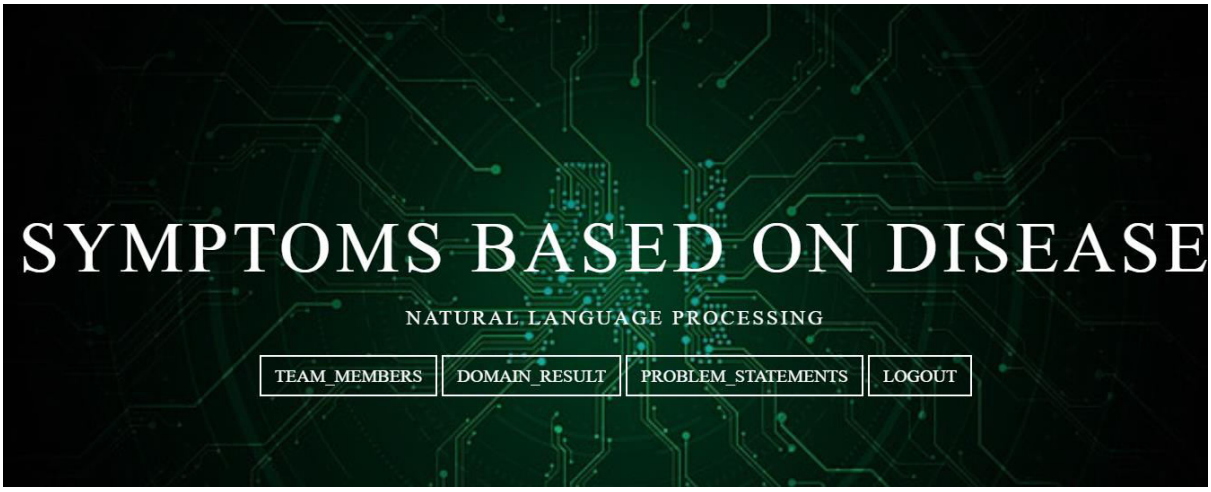
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

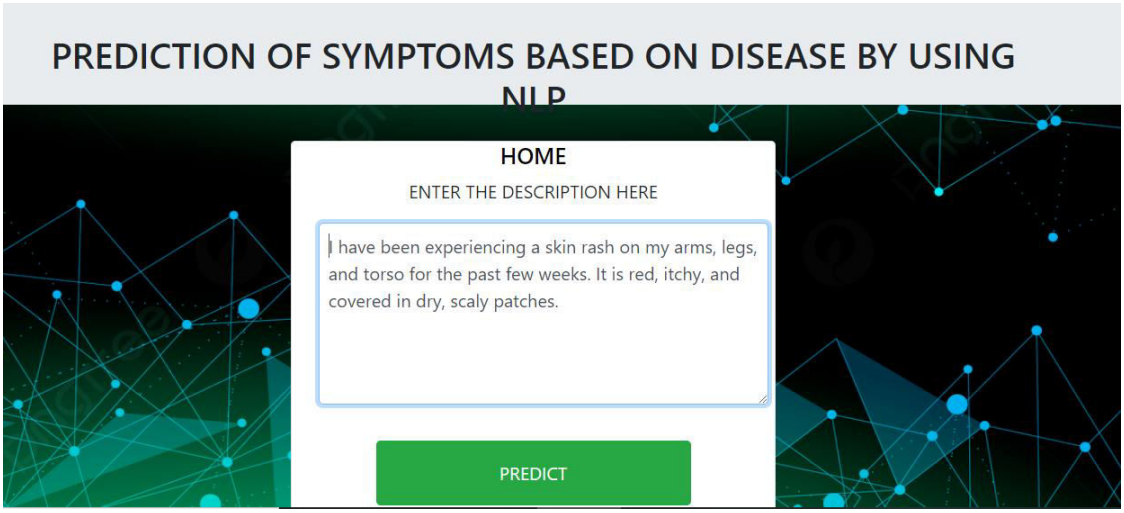
|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |

VI. RESULT





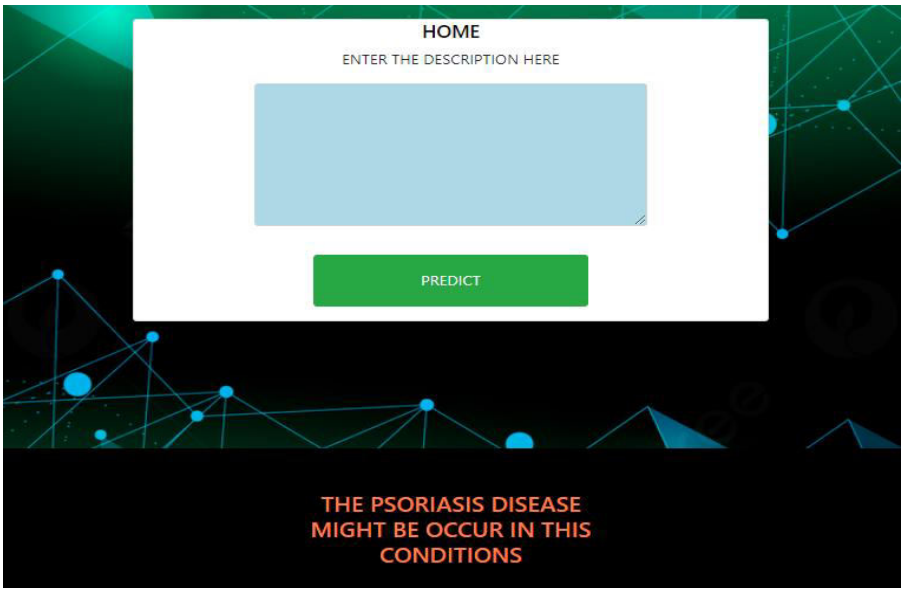


International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204359 |



VII. CONCLUSION AND FUTURE WORK

Employing Natural Language Processing (NLP) techniques for disease prediction based on symptoms holds tremendous potential for revolutionizing healthcare. The integration of advanced machine learning models and semantic analysis allows for more accurate and timely identification of diseases, facilitating early intervention and personalized medical care. Further research and development in this field could significantly enhance the efficiency of diagnostic processes, ultimately leading to improved patient outcomes and healthcare system effectiveness.

1.Future work can focus on refining disease prediction through NLP by developing advanced semantic embedding that capture subtle contextual relationships between symptoms, enhancing the model's ability to discern complex patterns in medical text data.

2.Explore the integration of continuous learning mechanisms to adapt and update the disease prediction model over time, accommodating new medical knowledge, evolving symptom patterns, and ensuring the system's relevance in dynamic healthcare environments.

REFERENCES

1. F. Pedregosa et al., ‘‘Scikit-learn: Machine learning in Python,’’ *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, Jul. 2017.
2. B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, ‘‘Librosa: Audio and music signal analysis in Python,’’ in *Proc. 14th Python Sci. Conf.*, vol. 8, pp. 18–25.
3. A. Paszke, S. Gross, and, ‘‘Pytorch: An imperative style, high-performance deep learning library,’’ in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.
4. F. Eyben, M. Wöllmer, and B. Schuller, ‘‘Opensmile: The Munich versatile and fast open-source audio feature extractor,’’ in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1459–1462.
5. H. B. Mann and D. R. Whitney, ‘‘On a test of whether one of two random variables is stochastically larger than the other,’’ *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.
6. X. Sun and W. Xu, ‘‘Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves,’’ *IEEE Signal Process. Lett.*, vol. 21, no. 11, pp. 1389–1393, Nov. 2014.
7. S. Kumar, T. S. Thambiraja, K. Karuppanan, and G. Subramaniam, ‘‘Omicron and Delta variant of SARS-CoV-2: A comparative computational study of spike protein,’’ *J. Med. Virol.*, vol. 94, no. 4, pp. 1641–1649, 2022.
8. D. Bhattacharya et al., ‘‘Analyzing the impact of SARS-CoV-2 variants on respiratory sound signals,’’ in *Proc. Interspeech*, Sep. 2022, pp. 2473–2477.
9. F. Avila et al., ‘‘Investigating feature selection and explainability for COVID-19 diagnostics from cough sounds,’’ in *Proc. Interspeech*, *IEEE Int.*

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details