iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51566

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Survival Study on Medical Disease Diagnosis with Coronavirus Database

S. Sathish Kumar[1], Dr. P. Parameswari[2]
[1]*Research Scholar M.C.A, Department of Computer Science, Karuppannan Mariappan College, Muthur Thirupur (Dt), Pin-638105, Tamilnadu, India*
[2]*Principal M. Sc., M.Phil., Ph.D., Palanisamy College of Arts No. 17, Erode road, Perundurai, Pin-638052. Erode (Dt), Tamilnadu, India*

Abstract: ***Data mining is a new technology with enormous potential to focus on important information and collected customer behaviour. Healthcare services used to have incredible potential for data mining because of this exponential development in the number of electronic health records. Medical data is health-related data related to regular patient care and clinical trial program. The feature selection process is choosing the relevant features for efficient disease diagnosis. With healthcare data, the classification and clustering methods had used for efficient disease diagnosis. In this paper, several researchers researched different disease diagnosis methods. However, time complexity had not reduced, and existing disease diagnosis methods did not enhance the accuracy. To address the proposed methods and different disease diagnosis schemes appraised.***
Keywords: Data mining, medical data, classification, clustering, disease diagnosis, healthcare data

I. INTRODUCTION

Data mining is a crucial area of research used in diverse areas such as finance, clinical research, education and healthcare. ML technique is used in the healthcare domain to diagnose different diseases. Medical data stored healthcare information such as patient records. Medical data mining has aimed at different data mining methods in medical applications. A medical data repository collects medical data stored and is accessible in different formats. Medical data mining is an important problem in obtaining valuable clinical knowledge from the medical database. Classification is the method of guaranteeing that unclassified data has included in particular categories. The paper has six sections: Section II reviews disease diagnosis techniques. Section III describes the existing disease diagnosis techniques. Section IV explains the simulation settings with the possible comparison between them. Section V discusses the limitation of existing disease diagnosis techniques. Finally, section VI concludes the paper.

II. LITERATURE SURVEY

In [1], a highly-scalable and robust machine learning framework had performed to forecast the adversity through mortality and ICU admission from vital sign time series. However, it failed to minimise the computational complexity. A straightforward clinical data encoding model into fixed-length feature vector representation had carried out in [2] with efficient feature selection from representation. However, the accuracy level was not improved. An MPCA had designed in [3] to minimise the bulk data. The minimised data was specified to classify the disease with better accuracy. As a result, the accuracy level had enhanced, and the computational cost had not reduced. Furthermore, a deep ensemble framework had employed in [4] for early prediction of COVID-19 from respective chest X-rays of patients. As a result, MPCA achieved better accuracy on the test sample with minimum false optimistic prediction. Moreover, prediction accuracy had not improved by the deep ensemble framework. In [5], a deep Spatio-temporal meta-learning model had designed to predict the traffic revitalisation index (DeepMeta-TRI) with external auxiliary information. However, the computational cost was not decreased by DeepMeta-TRI. A flexible and unsupervised data-driven approach was designed in [6] to identify the coronavirus infection depending on blood test samples. The designed method obtained the feature extraction ability of VAE and the detection sensitivity of the SVM algorithm. However, it failed to enhance accuracy with a flexible and unsupervised data-driven approach. In [7], an unsupervised machine learning method had used to stratify the patients with multiple sclerosis (pwMS) depending on brain MRI-derived volumetric features. However, the computational complexity level had not minimised. The SRMA technique had performed in [8] to a fully-labelled source dataset. SRMA reassigned the discriminative data using labelled source data into the target domain without any tissue annotation. However, the time complexity had not minimised by SRMA. In [9], depending on deep learning without competing for risks performed, the semi-supervised multitask learning (SSMTL) method.

The designed method converted to survival analysis with multitask learning by semisupervised learning and multipoint survival probability prediction. However, time consumption had not reduced by the SSMTL method. A DCAE method was determined in [10] to recover the clusters of chronic cough patients depending on data from the Electronic Medical Records (EMR) system. However, computational complexity had not minimised by DCAE. Therefore, a new algorithmic pipeline method was performed in [11] to automate the detection of unsupervised standardised gait tests using continuous IMU data. However, the accuracy level had not reduced by the algorithmic pipeline.

III. MEDICAL DISEASE DIAGNOSIS

The medical diagnosis method is an identifying disease or situation to explain the person's symptoms and signs. The data acquired for diagnosis had collected by history and physical examination of the person seeking medical care. Diagnostic events like medical tests had performed in the disease identification process. The new virus of COVID-19 had a pandemic in the world from the end of 2019. The new virus spreads much more with infection rate is improved. Governments took several drastic measures to handle the COVID-19 infection spread and residents' quarantine. Various efforts had taken to mitigate and delay the COVID-19 transmission for different applications likes

1) Wearing mask detection
2) COVID-19 spread forecasting
3) Chest X-ray diagnosis

A. A Knowledge Distillation Ensemble Framework for Predicting Short and Long-term Hospitalisation Outcomes from Electronic Health Records Data

Predict the adversity of mortality and ICU admission using a highly-scalable and robust machine learning method. Since then, readmission using a time series of vital signs and laboratory results determined within 24 hours of hospital admission. The two components of the stacked ensemble platform comprised the unsupervised LSTM Autoencoder and gradient boosting model. An unsupervised LSTM Autoencoder learnt the optimal version of time-series to distinguish minus frequent patterns that conclude by an adverse event from majority patterns—a gradient-boosting model designed to purify the prediction while incorporating static features. The designed methods appraised the patient's adversity risk and offered visual justifications. An ensemble Machine Learning framework termed Knowledge Distillation Outcome Predictor (KD-OP) addresses current difficulties in predicting adverse clinical outcomes using electronic health records data—the two learner modules of the designed framework comprised. Namely, Dynamic-KD learnt multivariate time series of patient physiology, and Static-OP estimated the adversity risk with static features. KD-OP employed the stacked architecture to collect the interplay with two patient views by Dynamic-KD to guide predictions during Static-OP. KD-OP had performed to befit the relative infrequency of adverse results in accurate hospital data.

B. Efficient Analysis Of Covid-19 Clinical Data Using Machine Learning Models

Extensive data had used to remove the relevant data into existence of variety and several levels of veracity. It was an essential one for COVID-19 future pandemics. A fixed-length feature vector representation had performed into the straightforward encoding of clinical data. The designed model performed efficient feature selection from representation. COVID-19 patients used a machine learning algorithm downstream for classification. An efficient feature selection algorithm attained higher prediction accuracy. The straightforward encoding helped the policymakers with specific attributes aiming at multiple random factors. ML models obtain the fixed-length feature vectors to implement classification and clustering. Fixed-length feature vector representation splits into clinical data attributes. Feature selection increased the predictive performance of classifiers and reduced the training runtime.

C. Multilinear Principal Component Analysis with SVM for Disease Diagnosis on Big Data

Map-reducing process and classification introduced an improved rule mining technique. An input of medical data had transferred to the map-reduce framework. MPCA had determined to reduce the bulk data. The reduced data was transferred to the classification process, and classified the disease with improved accuracy. To improve the accuracy and error rate minimised, the SVM classifier. Improved MPCA is a generalisation capability in the image reconstruction process. Map reduce framework in their data had rearranged into a 3D tensor. The SVM classifier was employed to classify and identify the disease by the resultant decreased data.

SVM was a two-class classifier with a hyperplane for categorising two data segments. The data point subset described the hyperplane location called support vectors.

D. Early prediction of COVID-19 using an ensemble of transfer learning

Automated disease detection considers an essential part of medical science used for the infectious nature of the coronavirus. A deep ensemble framework of transfer learning model had performed by advanced COVID-19 prediction from respective patients' chest X-ray images. Dataset gathered from Kaggle repository with COVID-19 Positive class and COVID-19 Negative class. The designed framework achieved improved accuracy on the test sample with minimum false optimistic prediction. The designed framework is an aid to doctors and technicians with early COVID-19 infection detection. The patient's health was observed distantly with connected devices and IoMT. IoMT-based solution for automatic COVID-19 detection was a significant step in pandemic fighting.

E. A Deep Spatio-Temporal Meta-Learning Model For Urban Traffic Revitalisation Index Prediction In The Covid-19 Pandemic

The COVID-19 pandemic is a global public health issue which caused adversity to people's average production and life. City managers use the Traffic Revitalisation Index Prediction Method to devise policies based on traffic and epidemic prevention. A deep Spatio-temporal meta-learning model had performed to predict the traffic revitalisation index (DeepMeta-TRI) with external auxiliary information like COVID-19 data. In deep Spatio-temporal meta-learning, meta-learning and external auxiliary information were united towards predicting the urban traffic revitalisation index. In order to do this, two models of Meta graph convolution network and Meta temporal convolution had performed to diverse Spatio-temporal correlations. The feature extraction of TRI context had processed by the Meta gating fusion module. Meta learners, respectively, create the parameter weights.

F. Deep Generative Learning-Based 1-SVM Detectors for Unsupervised COVID-19 Infection Detection Using Blood Tests

A flexible and unsupervised data-driven approach had introduced to identify the COVID-19 infection with the help of the blood test samples. The COVID-19 infection detection issues had recognised with a blood test as anomaly detection problems and an unsupervised deep hybrid model. The feature extraction capabilities of VAE and the detection sensitivity of the 1SVM algorithm had united by the designed method. The routine blood test samples were determined to measure the performance of deep learning models. The deep learning-driven 1SVM detection approach attained better detection performance.

IV. PERFORMANCE ANALYSIS OF DIFFERENT DISEASE DIAGNOSIS TECHNIQUES

In order to determine the different disease diagnosis techniques, input to conduct the simulation work took the amount of data points. The experiment used COVID-19 India Datasets, whose URL is https://www.kaggle.com/datasets/n1sarg/covid19-india-datasets?select=covid19india.csv. The number of new cases is rising day by day around the world. Since then, the dataset has information on India's states and union territories of India at a daily level. Experimental evaluation of six methods, namely highly-scalable and robust machine learning, straightforward encoding, improved rule mining technique, novel algorithmic pipeline, unsupervised machine learning and flexible and unsupervised data-driven approach, are carried out. Result analyses of existing disease diagnosis techniques had estimated with specific parameters: Disease Diagnosis accuracy, Disease Diagnosis time and Error rate.

A. Disease Diagnosis Accuracy

DDA is the ratio of the number of patient data points to diagnose to the total number of data points accurately. It had measured in terms of percentage (%). Disease diagnosis accuracy had computed as,

$$DDA(\%) = \frac{Number\ of\ disease\ data\ points\ that\ are\ correctly\ diagnosed}{N} * 100 \qquad (1)$$

From (1), the disease diagnosis accuracy had determined. 'N' denotes the number of data points

Table 1 Tabulation of Disease Diagnosis Accuracy

Number of Data points (Number)	Disease Diagnosis Accuracy (%)					
	Highly-scalable and robust machine learning	Straight forward encoding	Improved rule mining technique	Deep ensemble framework	DeepMeta-TRI	Flexible and unsupervised data-driven approach
100	94	85	78	80	75	70
200	95	88	80	82	77	72
300	93	86	76	79	76	71
400	92	85	74	76	73	69
500	91	82	71	74	70	67
600	89	80	69	72	68	65
700	90	81	68	70	66	63
800	92	83	70	73	67	66
900	93	85	72	75	70	68
1000	94	88	75	78	72	70

Table 1 above explains the disease diagnosis accuracy concerning the number of patient data points varying from 100 to 1000. When the number of data packets increases, the disease diagnosis accuracy increases or decreases respectively. Let us consider the number of data packets is 400, disease diagnosis accuracy of highly-scalable and robust machine learning, straightforward encoding, improved rule mining technique, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach is 92%, 85%, 74%,76%, 73% and 69% respectively. The graphical representation of disease diagnosis accuracy had described in Figure 1.

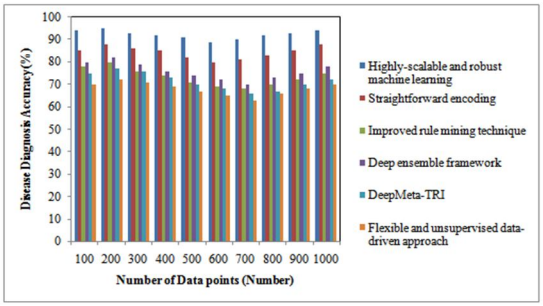


Figure 1 Measurement of Disease Diagnosis Accuracy

Figure 1 represents the disease diagnosis accuracy measure versus the number of data packets. Since the figure, the disease diagnosis accuracy of highly-scalable and robust machine learning is comparatively higher than other existing techniques. Due to the application of KD-OP for addressing the current problems for forecasting the difficult clinical results from electronic health records data. In this way, the disease diagnosis accuracy gets improved. As a result, highly-scalable and robust machine learning increases disease diagnosis accuracy by 10%, 26%, 22%, 30%, and 36% compared to straightforward encoding, improved rule mining technique, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach respectively.

B. Disease Diagnosis Time

DDT is the time consumed to diagnose the disease with patient data. It had measured in terms of milliseconds. It had formulated as,

$$DDT = N * Time\ consumed\ to\ diagnose\ one\ patient\ dat \quad (2)$$

From (2), the disease diagnosis time 'DDT' is determined. 'N' denotes the number of patient data points.

Table 2 Tabulation of Disease Diagnosis Time

Number of Data points (Number)	Disease Diagnosis Time (ms)					
	Highly-scalable and robust machine learning	Straightforward encoding	Improved rule mining technique	Deep ensemble framework	DeepMeta-TRI	Flexible and unsupervised data-driven approach
100	29	21	33	37	42	45
200	32	23	35	39	44	48
300	34	25	38	42	46	50
400	37	28	41	45	49	52
500	40	31	43	48	53	55
600	42	33	46	50	56	58
700	44	36	49	52	59	61
800	47	39	51	55	62	63
900	49	42	54	58	65	66
1000	52	45	57	60	68	70

Table 2 describes the disease diagnosis time concerning the number of patient data points varying from 100 to 1000. When the number of data packets increases, disease diagnosis time improves. It considers the number of data packets is 700, disease diagnosis time of highly-scalable and robust machine learning, straightforward encoding, improved rule mining technique, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach is 44ms, 36ms, 49ms, 52ms, 59ms and 61ms correspondingly.

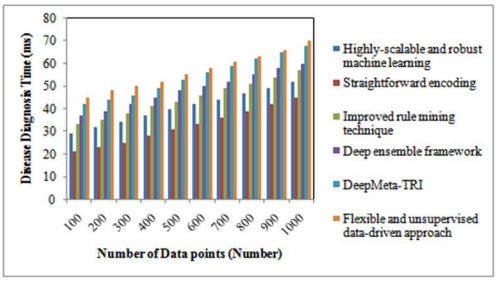


Figure 2 Measurement of Disease Diagnosis Time

Figure 2 explains the disease diagnosis time measure versus the number of data packets. In Figure 2, the disease diagnosis time of straightforward encoding is comparatively shorter than other existing techniques. On the contrary, experimental representation applies efficient feature selection. For example, COVID-19 patients employed machine learning for classification purposes. As a result, the feature selection algorithm achieved better accuracy with reduced time consumption. As a result, straightforward encoding consumes lesser disease diagnosis time by 21%, 29%, 34%, 41%, and 44% when compared to the highly scalable and robust machine learning, improved rule mining technique, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach respectively.

C. Error Rate

The error rate is the percentage of the number of patient data points wrongly diagnosed to the total amount of data points. It had measured in terms of percentage (%). The error rate had formulated as,

$$ER(\%) = \frac{Number\ of\ disease\ data\ points\ that\ are\ incorrectly\ diagnosed}{N} * 100 \quad (3)$$

From (3), the error rate had computed. 'N' is several data points, and ER is Error Rate.

Table 3 Tabulation of Error Rate

Number of Data points (Number)	Error Rate (%)					
	Highly-scalable and robust machine learning	Straightforward encoding	Improved rule mining technique	Deep ensemble framework	DeepMeta-TRI	Flexible and unsupervised data-driven approach
100	40	42	21	25	30	35
200	38	40	19	28	32	37
300	36	38	17	26	29	34
400	39	41	20	29	33	36
500	36	39	18	27	31	33
600	35	37	17	25	28	30
700	33	35	15	22	26	28
800	36	38	19	26	29	31
900	38	40	21	30	32	34
1000	40	42	23	33	35	37

Table 3 describes the error rate concerning the number of patient data points varied from 100 to 1000. When the number of data packets improved, the error rate had reduced or minimised, respectively. Let us consider the number of data packets is 600, the error rate of highly-scalable and robust machine learning, straightforward encoding, improved rule mining technique, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach is 35%, 37%, 17%, 25%, 28% and 30% respectively.

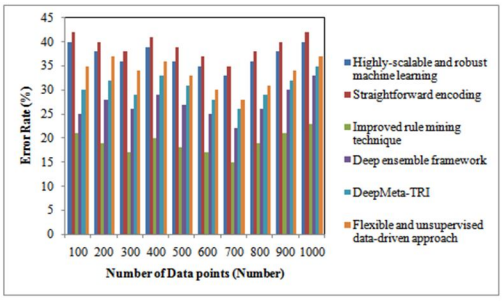


Figure 3 Measurement of Error Rate

Figure 3 explains the error rate measure versus the number of data packets. Since the figure, the error rate of the developed rule mining technique is comparatively lesser than other existing techniques. MPCA for minimising bulk data. The reduced data was transmitted to the classification process to classify the disease accurately. SVM classifier improved the accuracy and reduced the error rate. As a result, improved rule mining technique attains lesser error rate by 49%, 52%, 30%, 38%, and 43% when compared to highly-scalable and robust machine learning, straightforward encoding, deep ensemble framework, DeepMeta-TRI and flexible and unsupervised data-driven approach respectively.

V. DISCUSSION OF LIMITATIONS ON DISEASE DIAGNOSIS TECHNIQUES

A highly scalable and robust machine learning method had performed to forecast the adversity denoted by mortality and ICU admission with readmission from vital sign time series. An unsupervised LSTM Autoencoder studied an optimal symbol of time series to decide less frequent patterns with the adverse event from majority patterns. Based on learned temporal and static context, the stacked architecture created the visual justification of its predictions. However, computational complexity had not reduced by the designed method. The straightforward encoding allowed the policymakers to perform the task with attributes on multiple random factors. Fixed-length feature vector representation comprised the number of clinical data attributes. However, the accuracy level had not developed by the designed encoding model.

MPCA was engaged in array analysis, where a cube or hypercube of numbers is called a data tensor. SVM was employed to find an optimal margin and determined through the least distance among hyperplanes. As a result, the accuracy level had enhanced, and it failed to minimise the computational cost—a deep ensemble framework designed for early COVID-19 prediction. Designed framework helped the doctors and technicians with early COVID-19 infection detection. Unfortunately, though, the accuracy level was not improved.

DeepMeta-TRI employed with external auxiliary information. Two convolution network modules had employed for diverse spatiotemporal correlations. However, the computational cost was not decreased by DeepMeta-TRI. A flexible and unsupervised data-driven approach had determined to identify the COVID-19 infection with blood test samples. The flexible and unsupervised data-driven approach did not improve the accuracy level

A. Perceptions

An efficient disease diagnosis using deep learning and machine learning methods with improved accuracy and minimum time consumption carried the perception of the proposed work.

VI. CONCLUSION

A relative analysis of several disease diagnosis techniques had performed. The study failed to improve the disease diagnosis accuracy with a flexible and unsupervised data-driven approach. In addition, the disease diagnosis time had not reduced. The error rate had not reduced. The comprehensive experiment on conventional techniques estimates the result of different disease diagnosis techniques and discusses its issues. From the result analysis, the research had carried out using machine learning, and ensemble learning techniques had designed for efficient disease diagnosis with improved accuracy and minimum time consumption.

REFERENCES

[1] Zina M. Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, James T. H. Teo, and Richard JB Dobson, "A Knowledge Distillation Ensemble Framework for Predicting Short- and Long-Term Hospitalization Outcomes From Electronic Health Records Data", IEEE Journal of Biomedical and Health Informatics, Volume 26, Issue 1, January 2022, Pages 423-435

[2] Sarwan Ali, Yijing Zhou, and Murray Patterson, "Efficient analysis of COVID-19 clinical data using machine learning models", Medical, Biological Engineering & Computing, Springer, Volume 60, 2022, Pages 1881–1896

[3] Juby Mathew and R. Vijaya Kumar, "Multilinear Principal Component Analysis with SVM for Disease Diagnosis on Big Data", IETE Journal of Research, Volume 68, Issue 1, 2022, Pages 526-540

[4] Pradeep Kumar Roy and Abhinav Kumar, "Early prediction of COVID-19 using an ensemble of transfer learning", Computers and Electrical Engineering, Volume 101, July 2022, Pages 1-15

[5] Yue Wang, Zhiqiang Lv, Zhaoyu Sheng, Haokai Sun and Aite Zhao, "A deep Spatio-temporal meta-learning model for urban traffic revitalisation index prediction in the COVID-19 pandemic", Advanced Engineering Informatics, Elsevier, Volume 53, August 2022, Pages 1-12

[6] Abdelkader Dairi, Fouzi Harrou, Member, and Ying Sun, “Deep Generative Learning-Based 1-SVM Detectors for Unsupervised COVID-19 Infection Detection Using Blood Tests”, IEEE Transactions on Instrumentation and Measurement, Volume 71, 2022, Pages 1-11

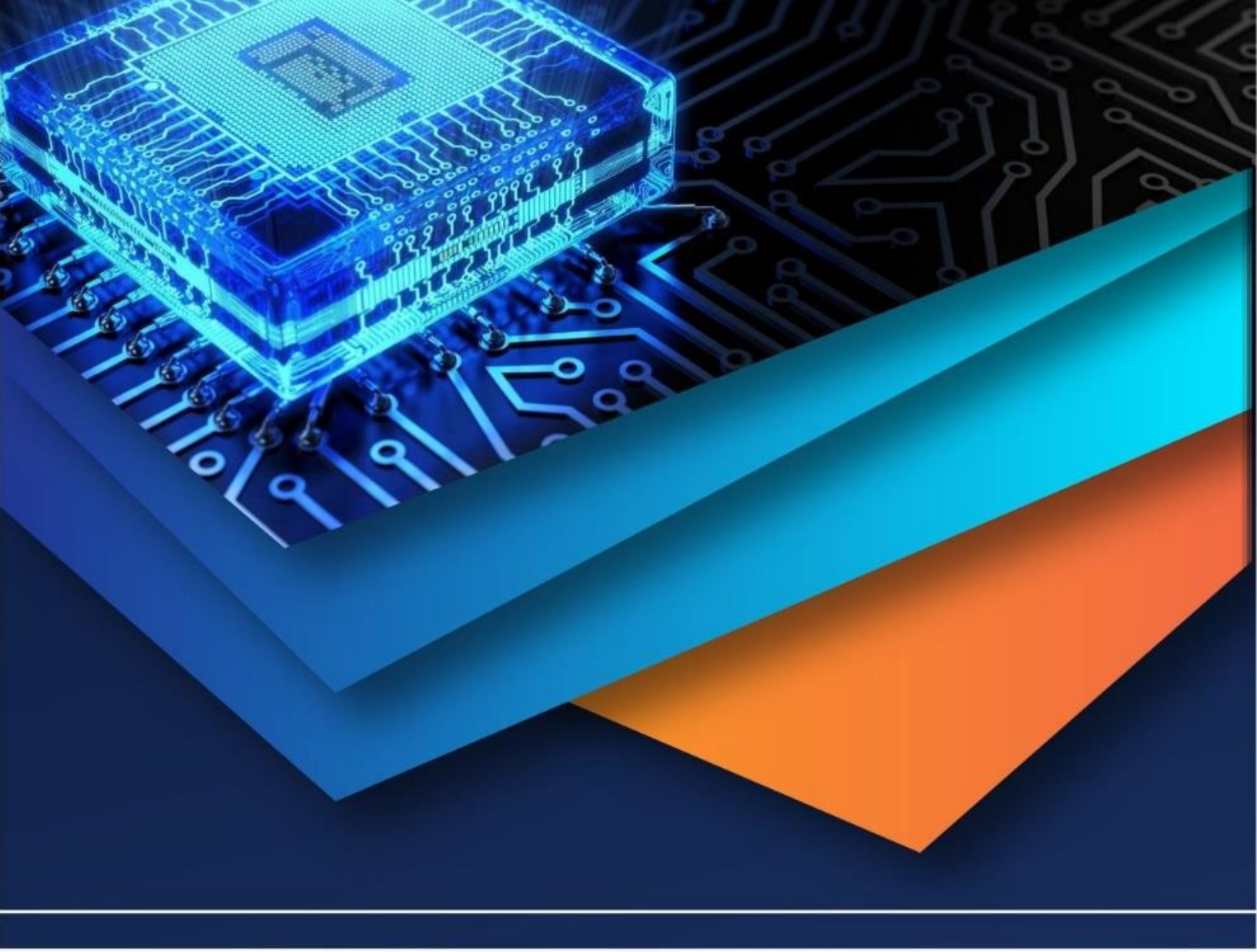
[7] Giuseppe Pontillo, Simone Penna, Sirio Cocozza, Mario Quarantelli, Michela Gravina, Roberta Lanzillo, Stefano Marrone, Teresa Costabile, Matilde Inglese, Vincenzo Brescia Morra and Daniele Ricci, Andrea Elefante, Maria Petracca, Carlo Sansone and Arturo Brunetti "Stratification of multiple sclerosis patients using unsupervised machine learning: a single-visit MRI-driven approach", European Radiology, Springer, Volume 32, 2022, Pages 5382–5391

[8] Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, Jean-Philippe Tirana, "Self-rule to multi-adapt: Generalised multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection", Medical Image Analysis, Volume 79, 2022, Pages 1-20

[9] Shengqiang Chi, Yu Tian, Feng Wang, Yu Wang, Ming Chen and Jingsong Li, "Deep Semisupervised Multitask Learning Model and its Interpretability for Survival Analysis", IEEE Journal of Biomedical and Health Informatics, Volume 25, Issue 8, August 2021, Pages 3185 – 3196

[10] Wei Shao, Xiao Luo, Zuoyi Zhang, Zhi Han, Vasu Chandrasekaran, Vladimir Turzhitsky, Vishal Bali, Anna R. Roberts, Megan Metzger, Jarod Baker, Carmen La Rosa, Jessica Weaver, Paul Dexter and Kun Huang, "Application of unsupervised deep learning algorithms for identification of specific clusters of chronic cough patients from EMR data", BMC Bioinformatics, Volume 23, Issue 140, 2022, Pages 1-14

[11] Martin Ullrich, Annika Mucke, Arne Kuderle, Nils Roth, Till Gladow, Heiko Gaßner, Franz Marxreiter, Jochen Klucken, Bjoern M. Eskofier and Felix Kluge, "Detection of Unsupervised Standardised Gait Tests From Real-World Inertial Sensor Data in Parkinson's Disease", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Volume 29, 2021, Pages 2103-2111



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY