iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60698

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 12 Issue IV Apr 2024- Available at www.ijraset.com

Multiple Disease Prediction using ML

Prof. A.D.Mhatre[1], P.R.Umale[2], A.P.Pawar[3], P. P.Mhatre[4], A.S.Singh[5]
Department of Information Technology, Datta Meghe College of Engineering, University of Mumbai

Abstract: ***This study proposes a multi-disease prediction system integrating assessments for diabetes, heart disease, and Parkinson's disease. The system leverages machine learning algorithms like Logistic Regression, Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN) to predict disease risk through a unified user interface. We delve into the system design process, emphasizing the importance of defining system compo- nents and exploring the utility of modeling languages. The research explores the potential of SVMs, including linear and non-linear variations, for disease prediction. We analyze existing literature on applying machine learning algorithms for disease prediction and discuss their potential for disease classification. Finally, the abstract addresses challenges and future directions in disease prediction, aiming to provide valuable insights for further research and development efforts.***

Keywords: Python, KNN, XGBoost, SVM, Logistic Regression, Diabetes, Parkinsons, Heart- attack

I. INTRODUCTION

Diabetes, heart disease, and Parkinson's disease stand as formidable contributors to mortality in con- temporary society. Heart disease, encompassing con- ditions like arrhythmias, coronary artery disease, and congenital defects, remains a leading cause of death in India. According to a 2021 study published in the Lancet journal, an estimated 2.8 million Indians died from cardiovascular diseases in 2019, accounting for 28In the realm of diabetes, characterized by elevated blood sugar levels, a growing threat looms in India. The International Diabetes Federation (IDF) reports that in 2023, an estimated 77 million adults in In- dia were living with diabetes, marking a staggering 18Parkinson's disease, while less prevalent than di- abetes and heart disease, presents a significant chal- lenge in India. The Movement Disorders Society esti- mates that approximately 5-10 lakh individuals in In- dia live with Parkinson's disease, a number expected to increase due to the aging population. Our research aims to redefine disease prediction by integrating di- abetes, heart disease, and Parkinson's disease into a unified system. Traditional medical AI models typ- ically focus on individual diseases in their examina- tions, lacking a cohesive framework for multi-disease predictions. In our proposed system, we leverage ma- chine learning classification algorithms such as Lo- gistic Regression, Support Vector Machine (SVM), and K-Nearest-Neighbors (KNN) to predict the on- set of diabetes, heart disease, and Parkinson's disease through a single user interface.

II. SYSTEM DESIGN

System design, in the context of developing com- puter systems, is the intricate process of defining the armature, factors, modules, interfaces, and data to align with specific criteria. This operation bridges systems proposition to product development. Object-oriented design and analysis styles have swiftly be- come the most favored methodologies for construct- ing computer systems in the realm of a burgeoning Indian dataset and population.

Description: System design involves meticulously defining system elements such as modules, arma- ture, factors, interfaces, and data, grounded in the conditions specified for the system. It is the sys- tematic process of defining, developing, and design- ing systems tailored to meet the specific require- ments and conditions of businesses or associations. A harmonious and well-performing system demands a methodical approach, necessitating a top-down or bottom-up approach to consider all pertinent system variables. Developers utilize modeling languages to express information and knowledge in a structured manner defined by cohesive rules and delineations. Plans can be articulated using visual or textual mod- eling languages.

In the context of the Indian dataset and popula- tion, some exemplary graphical modeling languages include:

1) Unified Modeling Language (UML), describ- ing software both structurally and functionally with graphical notation.
2) Flowchart, offering a schematic or step-by-step representation of an algorithm.
3) Business Process Modeling Notation (BPMN), utilized in a process modeling language.
4) Systems Modeling Language (SysML), applied for system design.

Design styles encompass:

a) Architectural design, detailing the views, mod- els, gestures, and structure of the system.

b) Logical design, representing the information in- flow, inputs, and operations of the system, with ex- amples like Entity-Relationship (ER) diagrams illus- trating reality relationship plates.

c) Physical design, encompassing how users input information into the system, how the system provides information back, how data is modeled and stored, and the flow, validation, defense, and conversion of data as it traverses through and out of the system.

A. Architectural Design

In the realm of software engineering, architectural de- sign involves the decomposition of the system into in- teracting factors. This process is depicted through a block illustration, providing an overview of the sys- tem's structure, the characteristics of its components, and the mechanisms through which these components interact to exchange data. Specifically tailored for developing computer-based systems, architectural de- sign identifies essential factors and delineates the re- lationships between them. It defines both the struc- ture and functionalities of the factors within the sys- tem, as well as the non-interactions between them.

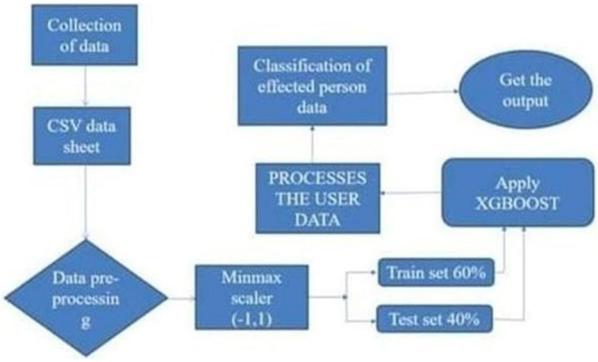


Fig. 1: Architectural design
Kothapeta et al. (2023)

The architectural design process revolves around es- tablishing connections between factors, also known as subsystems, that constitute the system. It out- lines the structure and interactions of these subsys- tems and their interplay. Serving as an early stage in the system design phase, architectural design acts as a crucial link between specification conditions and the subsequent design processes. This systematic ap- proach ensures that the design aligns seamlessly with the predefined criteria and requirements outlined in the earlier stages of the system development cycle.

B. Data flow model

Data inflow modeling proves valuable in identifying various effects, such as information input from or transfer to other entities, associations, or computer systems. It also encompasses the depiction of areas within a system where information is stored and the pathways through which information flows within the system. Additionally, the processes within a system that manipulate the entered information and execute various operations are effectively modeled through data inflow diagrams. This comprehensive approach ensures a thorough understanding of the data dynam- ics within the system and facilitates efficient design and development aligned with predefined criteria.

C. Support Vector Machine (SVM)

The Support Vector Machine (SVM) has emerged as a widely acclaimed supervised learning algorithm, originally renowned for its prowess in solving clas- sification and regression problems within the realm of machine learning. Historically associated with bracket problems, SVM has evolved into a versatile tool, finding applications in diverse fields such as pat- tern recognition, bioinformatics, and text classifica- tion. At its core, the SVM algorithm aims to estab- lish a discriminating decision boundary, known as a hyperplane, within an n-dimensional space. This hy- perplane serves the critical function of effectively seg- regating different classes, enabling the seamless clas-sification of new data points in the future.

Fig. 2: Data Flow MODEL

The mech- anism behind SVM involves the strategic selection of extreme point vectors, termed support vectors, which play a pivotal role in defining the hyperplane.

To enhance the efficiency of SVM training, par- ticularly in addressing the inherent quadratic opti- mization problem, the algorithm is often coupled with a Sequential Minimal Optimization (SMO) classifier. This approach, introduced by Platte, adds a layer of sophistication to the SVM classifier, contributing to its robust performance.

SVM manifests in two distinctive types, catering to the nature of the data:

1) Linear SVM: Applied when dealing with lin- early divisible data, where a single straight line can effectively classify data into two distinct classes. This type of SVM is appropriately named a linear SVM classifier.
2) Non-linear SVM: Employed when faced with non-linearly distributed data, instances where a straight line fails to provide an effective classifica- tion. In such scenarios, a non-linear SVM classifier comes into play, offering more flexibility in capturing complex relationships within the data.

Hyperplane: In the n-dimensional space, where multiple decision boundaries can segregate classes, the quest is for the optimal decision boundary – the SVM hyperplane. The configuration of this hyper- plane is contingent upon the features present in the dataset, presenting itself as a straight line for two fea- tures or a two-dimensional plane for three features.

Support Vectors Crucial to the SVM process, sup- port vectors are those data points situated closest to the hyperplane or vectors that significantly influence the position of the hyperplane. Serving as the foun- dational pillars supporting the hyperplane, these vec- tors are instrumental in the SVM classification pro- cess, ensuring accurate and robust outcomes.

D. Linear SVM

LINEAR SVM:

Linear SVM is a classification algorithm specifi- cally tailored for linearly divisible data. This implies that if a dataset can be effectively separated into two classes using a single straight line, it is considered lin- early divisible. In such cases, the classifier employed is aptly named a linear SVM classifier. The appli- cation of Linear SVM is particularly relevant when dealing with scenarios in the Indian dataset and pop- ulation where a clear linear boundary can distinctly categorize data.

To grasp the workings of the Linear SVM algo-rithm, let's consider an illustrative example. Imag- ine we have a dataset with two distinct markers, say green and blue, and the dataset is characterized by two features, denoted as x1 and x2. The objective is to design a classifier capable of categorizing pairs of coordinates (x1, x2) as either green or blue.

![Linear SVM Illustration](imagepath)

In the context of the Indian dataset, where lin- ear separability may be observed in certain as- pects, Linear SVM proves to be a valuable tool for classification tasks. Whether it's discerning pat- terns in health data, economic indicators, or other population-related features, the linear SVM classifier can efficiently draw a single straight line to distin- guish between different classes. This becomes par- ticularly relevant when dealing with diverse datasets representing the intricacies of the Indian population.

E. Non-Linear SVM

Non-linear SVM is a pivotal tool when dealing with non-linearly distributed data, a scenario often en- countered in diverse datasets, including those of the Indian population. This designation arises when a set of data points cannot be effectively classified using a straight line – the hallmark of linearly distributed data. In the context of Indian datasets, which often exhibit complex relationships and intricate patterns, non-linear SVM becomes particularly relevant.

In the traditional setting of linearly distributed data, a direct SVM classifier suffices, efficiently seg- regating data points with a straight line. However, for non-linear datasets prevalent in the Indian con- text, a direct SVM classifier may fall short of cap- turing the inherent complexities. Even when it may seem possible to separate such non-linear data with a straight line, the effectiveness of the classification process can be significantly enhanced by introducing an additional dimension.

Considering the Indian dataset, this additional di- mension is akin to addressing the intricacies and nu- ances present in the population characteristics.

By extending from the two-dimensional space (repre- sented by x and y in the linear scenario), a third dimension (z) is introduced to better encapsulate the complexity of the non-linear data. This additional dimension can be computed using the formula: Z = (x2) + (y2).

Illustrated in Fig-5, the Non-Linear Support Vec- tor Machine employs this third dimension, transform- ing the sample space. This transformation allows the SVM algorithm to effectively divide the datasets into classes, offering a more accurate representation of the complex relationships within the Indian dataset and population. The versatility of non-linear SVM makes it a valuable asset in deciphering intricate patterns and relationships, making it well-suited for address- ing the multifaceted nature of data prevalent in the Indian context.

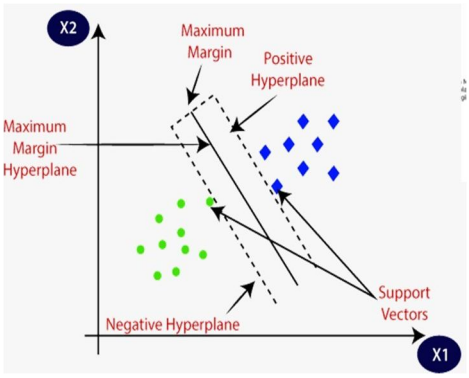


Fig. 3: Support vector Machine

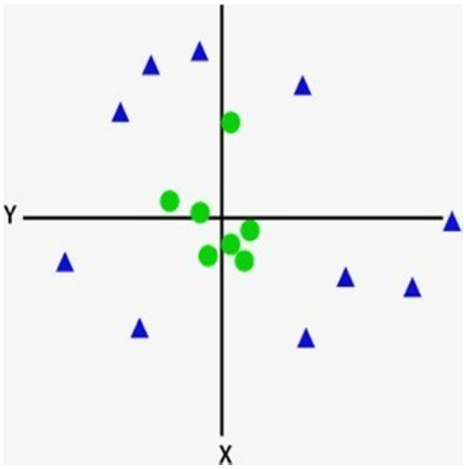


Fig. 4: Linear SVM

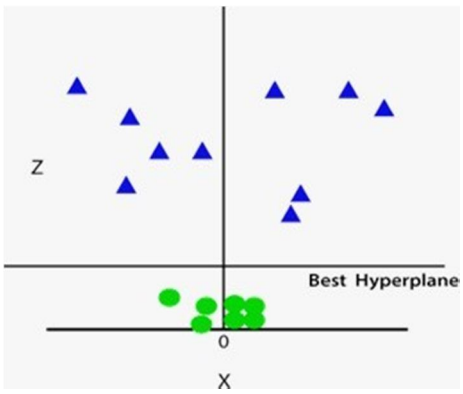


Fig. 5: Non Linear SVM

F. Formulas

Mathematical, physical, chemical symbols and for- mulas should clearly be typed in the standard scien- tific notations. Sort mathematical, physical, chemi- cal ... formulas can be written within the text lines, while long ones can be written separately in new lines. The sort formulas are not numbered. The long formu- las are numbered consecutively with arabic numbers within parentheses on the right-hand side of each for- mula. Only cited long formulas are numbered. The long formulas can grammatically have dots or com- mas at their ends like in the correct sentences gram- mar. For example, single equation looks like

$$\mathbf{G}(\mathbf{k}, \omega) = \mathbf{G}_0(\mathbf{k}, \omega) + \mathbf{G}_0(\mathbf{k}, \omega)\Sigma(\mathbf{k}, \omega)\mathbf{G}(\mathbf{k}, \omega). \quad (1)$$

The system of equations looks like

$$G_0(\mathbf{k}, \omega) = \frac{1}{\omega - \varepsilon(\mathbf{k})}, \quad (2)$$

$$\Sigma(\mathbf{k}, \omega) = Un + \frac{U^2 n(1-n)}{\omega - (1-n)\varepsilon(\mathbf{k})}. \quad (3)$$

Long formulas in appendices must be numbered separately, and the formulas number must include a capital roman letter identifying the appendix, for ex- amples (A1), (A2), (A3), etc.

III. LITERATURE SURVEY

Recent research has focused on leveraging machine learning algorithms like logistic regression, SVM, ran- dom forest, and Extra Tree Classifier for disease prediction. While individual models have shown promise, there's a growing need for unified platforms predicting multiple diseases simultaneously. Exist- ing projects demonstrated this with their platform, integrating diverse health indicators to assess risks for heart disease, diabetes, and hypertension. Addi- tionally, Streamlit Python Module has gained trac- tion for developing user-friendly interfaces in health- care applications, enhancing accessibility and usabil- ity. Future directions include expanding predictive scopes and refining algorithms for higher accuracy, aiming to contribute to early detection and interven- tion, thereby reducing mortality rates and improving public health outcomes.K et al. (2023) This study ad- dresses the scarcity of unified systems for forecasting multiple diseases. By utilizing various classification algorithms like K-Nearest Neighbor, SVM, Decision Tree, Random Forest, Logistic Regression, and Gaus- sian naive Bayes, the research aims to identify the most accurate algorithm for predicting diseases such as diabetes, heart disease, chronic kidney disease, and cancer. Through validation and comparison of algo- rithmic accuracies using multiple datasets for each disease, the study endeavors to develop a web appli- cation capable of early disease detection and diagno- sis, potentially saving countless lives.Gopisetti et al. (2023)

This research investigates the potential of various machine-learning algorithms for disease classification. The authors conduct a comprehensive comparative study, evaluating nine state-of-the-art techniques on eight diverse disease datasets. This research is antic- ipated to provide valuable insights into the strengths and weaknesses of various machine learning tech- niques for disease classification tasks. By analyzing diverse datasets and performance metrics, the study aims to inform the selection of appropriate algo- rithms for specific disease prediction problems.Azar et al. (2018) This research investigates the poten- tial of machine learning for disease prediction. The authors explore three popular machine learning al- gorithms: Support Vector Machine (SVM), Naive Bayes, and Random Forest. Their findings demon- strate that Random Forest achieved the highest ac- curacy (87%) among the three models. Further opti- mization through hyperparameter tuning yielded an even better accuracy of 90%. This study highlights the importance of algorithm selection and tuning for effective disease prediction using machine learn- ing.Jovovic et al. (2023)

The prominent issues that arise in the realm of disease classification/prediction using ML methods include data heterogeneity, class imbalance, and in- terpretability. Heterogeneity of data can be under- stood as multifaceted nature of the healthcare data (demographic, clinical, and biomarker information, for example) that can be distracting while integrat- ing and processing the data of different modalities for making predictions on a disease. One problem is the class imbalance, when the classes of different dis- eases are unequally distributed, leads to inconsistent model performance and inaccurate predictions.Some of the avenues for advancing prediction accuracy and modalities are incorporating the optimization tech- niques, the use of feature selection method and in- tegrating multiple data modalities. Things such as optimization techniques like grid search and cross- validation finetuning model hyperparameters, can lead to better performance.

Then feature selection approaches, for instance genetic algorithms and re- cursive feature elimination, may be used to find out the relevant features and dimensionality reduction thereby improve the model's disease pattern captur- ing capability. Moreover, the incorporation of diverse data modalities, including physiological/bio-signals and electronic health records helped to get a holistic view of the disease patterns and make the probability of forecasting better.Parshant and Rathee (2023)
The successful case studies demonstrate the appli- cation of machine learning in predicting multiple dis- eases. The study applies different classification algo- rithms to three separate databases of diseases: breast cancer, heart disease, and diabetes. The feature se- lection for each dataset is achieved through backward modelling using the p-value test. The results of the study show that machine learning algorithms, such as Logistic Regression, Decision Tree, Random For- est, Support Vector Machine, and Adaptive Boosting, have been successful in predicting diseases with high accuracy. For example, the prediction accuracy of the proposed method reaches 87.1 percent heart disease detection using Logistic Regression, 85.71 percent in diabetes prediction using Support Vector Machine (linear kernel), and 98.57 percent using AdaBoost classifier for breast cancer detection.Kohli and Arora (2018) The applications of machine learning (ML) in multiple disease prediction, including heart disease, diabetes, and breast cancer, have shown promising results. ML algorithms like Naive Bayes, SVM, and decision trees enable quick and accurate diagnosis. Incorporating adaptivity, SVM enhances diagnostic methods. Hybrid models combining clustering and classification have been successful in predicting type 2 diabetes with high accuracy. These applications un- derscore ML's potential in enhancing diagnostic pro- cesses and aiding clinical decision-making.Arumugam et al. (2021)

IV. PROPOSED METHODOLOGY

This describes a promising approach to building a reliable system for predicting multiple diseases with machine learning. The method breaks down into sev- eral key stages to guarantee thorough data handling, model selection, assessment, and use.
Here's a breakdown of the key steps:
Data Gathering and Cleaning:
The first step involves collecting extensive datasets that contain important characteristics for each disease. This might involve gathering information from various sources including elec- tronic medical records, imaging reports, and lab results. Once collected, the data goes through preprocessing to address missing information, outliers, and inconsistencies. Techniques like data cleaning, normalization, and encoding cat- egorical variables are used to ensure data quality and uniformity.
Feature Engineering and Selection: Next, fea- ture engineering is applied to extract valuable insights from the data. This could involve cre- ating new features, transforming existing ones, or choosing relevant features using techniques like correlation analysis or the importance of fea- tures in tree-based models. For unstructured
textual data, natural language processing (NLP) techniques are used to extract relevant features and sentiment analysis helps capture information from the text.
Model Development and Evaluation: A vari- ety of machine learning algorithms well-suited for disease prediction tasks are chosen, such as logistic regression, support vector machines (SVM), decision trees, random forest, and en- semble methods like XG Boost.

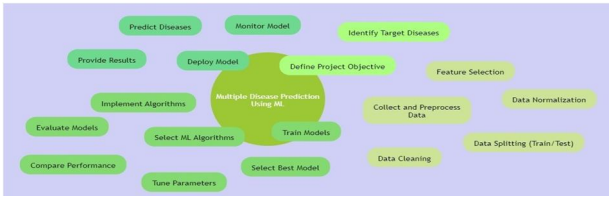


Fig. 6: Proposed Methodology

Each algorithm is trained on the preprocessed data and assessed using appropriate performance metrics like ac- curacy, precision, recall, and F1-score. Tech- niques like cross-validation are employed to eval- uate how well the models perform in general sit- uations and to avoid overfitting.
Model Interpretation and Validation: For mod- els like decision trees, the focus is on interpreting the decision-making process to gain insights into feature importance and model behavior. Visu- alization techniques are used to illustrate tree structures and decision boundaries. The trained models are validated on a separate test set toassess their performance in real-world scenarios.

Deployment and Integration: Finally, the mod- els with the best performance are deployed in a healthcare setting, where they are integrated into existing systems or applications and made accessible to healthcare professionals and end- users. Model deployment involves ensuring com- pliance with ethical and privacy regulations, as
well as collaborating with domain experts to vali- date the model's effectiveness in clinical practice.
By following this proposed methodology, the aim is to develop a powerful and scalable framework for pre- dicting multiple diseases with machine learning, ulti- mately contributing to improved healthcare outcomes and better patient care.Pattar et al. (2023)

V. CONCLUSIONS

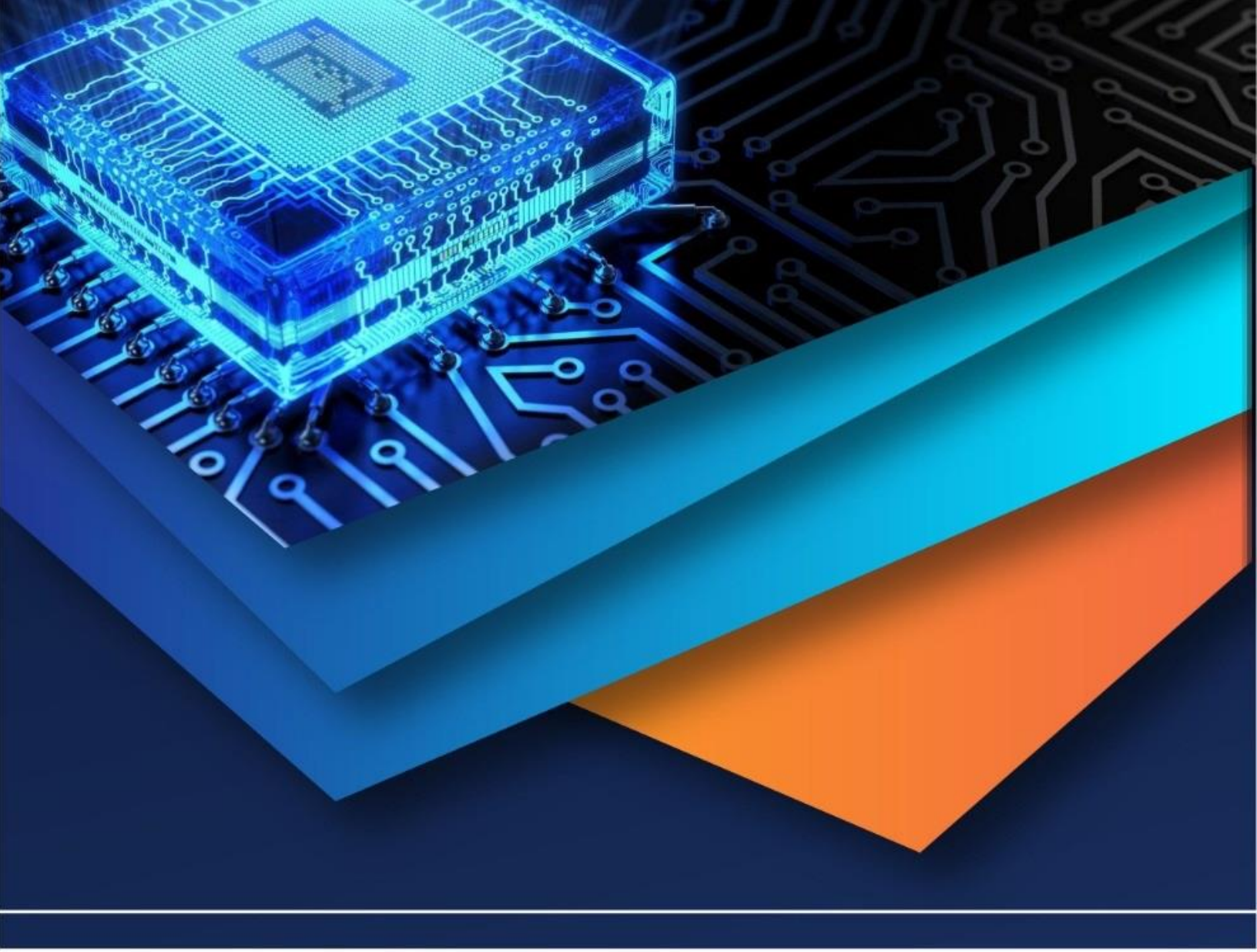
Here, we outlined a disease prediction model that identifies the possibility of diabetes, cardiovascular disorders and Parkinson's disease within a unified framework of the model. This system is different from the traditional medical AI models in which the diag- nosis of a single disease is aimed, but instead, it offers a more comprehensive one which is disease detection in the initial stages. Through utilizing machine learn- ing tools like Logistic Regression, SVM, and KNN our platform features an easily accessible interface that will help individuals to understand their possi- ble health threats for heart disease, stroke, and hyper- tension. Despite the current capacity, the potential of this research is promising to the future of preventa- tive healthcare. Detection of chronic diseases such as diabetes, heart disease and Parkinson's disease in the early stages is an important factor for disease control and treatment outcome. In this way, people have op- portunities to be active players on their own health through this system that delivers important risk as- sessments. Directions for future research are to add on the additional disease markers for extra accuracy and to check the use of deep learning algorithms for improvement in the performance of the system. As the result, the system would be a key for early disease detection and risk stratification for all different health problems, and can be a game changer in disease pre- diction in healthcare sphere that is more complete and more readily approachable.

VI. FUTURE SCOPE

Farther exploration could explore the possibility of combining a support vector machine algorithm to prize prophetic features with unsupervised literacy. likewise, only healthy cases were used in this analy- sis. A machine literacy model with analogous perfec- tion, delicacy and recall may be possible for a model that considers multiple affiliated judgments . More complex ML algorithms will be created in the future, much is demanded to ameliorate complaint predic- tion. In addition, learning models should be cali- brated more frequently after the training phase to achieve better results. In addition, datasets should be expanded to include different population groups to avoid over-fitting and ameliorate the data delicacy of applied models. Eventually, more important selec- tion styles should be used to ameliorate performance. In the future, we can add to the being complaint API. To reduce mortality, we can try to ameliorate the del- icacy of the cast. Try that the system is friendly and offers standard converse.

REFERENCES

[1] Arumugam, K., Naved, M., Shinde, P. P., et al. 2021, Materials Today: Proceedings,)
[2] Azar, D., Moussa, R., and Jreij, G. 2018, International journal of artificial intelligence, 16, 25
[3] Gopisetti, L. D., Kummera, S. K. L., Pattamsetti, S. R., et al. 2023, 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT),), 923
[4] Jovovic, I., Babic, D., Popovi´c, T., Cakic, S., and Katnic, 2023, 2023 27th International Conference on Infor- mation Technology (IT),), 1
[5] K, M., S, S., and E, T. 2023, 2023 International Confer- ence on Emerging Research in Computational Science (ICERCS),), 1
[6] Kohli, P. S. and Arora, S. 2018, 2018 4th International Conference on Computing Communication and Au- tomation (ICCCA),), 1
[7] Kothapeta, H., Lakkampelly, S., Mandari, A., Sathya- narayana, M., and Vani, G. 2023, Computer Sci-ence Sreenidhi Institute Of Science And Technology (SNIST),)
[8] Parshant and Rathee, D. A. 2023, Iconic Research And Engineering Journals, 6, 411
[9] Pattar, A., Kurhade, S., Salunke, M., Satav, D., and
[10] Priyadarshni, P. A. 2023, International Journal for Re- search in Applied Science and Engineering Technology,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY