ISSN:2231-0711

Soumya Pawar et al | IJCSET(www.ijcset.net) | 2021 | Volume 11, 31-35

# Discovering Pathological Disorders in Children through Artificial Neural Networks

**Soumya Pawar[#1], Mayuri Khandade[*2], Sayli Shinde[#3], Deepak Tamhane**

*Information Technology Department ,Modern College of Engineering*
*1186/A, Off J.M. Road, Shivaji nagar, Pune, Maharashtr, India 411005*

1soumyav.pawar@moderncoe.edu.in

3sayali.shinde@moderncoe.edu.in

2mayuri.khandade@moderncoe.edu.in

**Abstract – There has been an increase in the number of patients suffering from pathological disorders. Pathological Disorders are the type of disorders that do not belong to a particular category of diseases. These disorders exhibit symptoms that cannot be correlated to the conventional diseases or are correlating between two different types of diseases. There has been increased development in the healthcare sector where an improvement in the diagnosis and treatment of the patients is well observed. For improving the detection of the pathological disorders, an effective approach has been defined in this research article. This approach performs the pathological disorder classification through the use of machine learning algorithms for improved accuracy. The achieved methodology utilizes K Nearest neighbor along with Artificial Neural Networks and a combination of two different classification algorithms, such as Decision Tree and Random Forest Classification. The approach has been experimented on using extensive experimentation to achieve extremely positive results.**

**Keywords:K-Nearest Neighbor, Artificial Neural Network, Decision Tree, Random Forest Classification.**

## I. INTRODUCTION

Humans have been noticing immense growth and advancement since the development of language. Before the development of language, humans use to communicate with each other through the use of sign language and different sounds. The early humans refined the language more and more and develop effective and useful communication techniques to collaborate with one another. This has allowed greater cooperation between various human beings which has been effective in catalyzing growth and achieving significant milestones. This would not have been possible without the development of an effective communication system in the language.

The development and improvement of the language have helped in achieving significant goals. Communication has allowed humans to change from their Hunter-gatherer lifestyle to a more settled lifestyle by being able to grow crops by collaborating with one another. Even hunter-gatherers utilized primitive language to bring down large prey in packs.

Therefore it can be understood that language plays a very effective and significant role in human being's life. This can also be seen nowadays where increasing amounts of communication and faster modes of communication are preferred largely by the population.

This hunger for knowledge and information for the purpose of communicating this information has been vital to the development and advancement of various technologies all over the globe. This has also improved the approaches in the medical paradigm which has led to significant improvement in the treatment and diagnosis of the ailments. The improvements in the medical sector have proven to be considered as it has largely effective in the reduction of mortality rates all over the globe.

Due to the increasing population, there has been a significant increase in the number of patients. This causes a shortage of doctors and medical professionals who have to manage their time and provide it to the needy person with an urgent medical emergency. These results in a large number of patients not being diagnosed which leads to lower satisfaction and higher incidences of diseases especially pathological disorders. This is due to the fact that pathological disorders are a highly difficult and complex mechanism that is difficult to detect and identify.

Most of the time the doctors are also unable to identify the pathological disorders effectively due to their inherent complications and complexity. Therefore there is a need for an effective mechanism that can accurately identify the pathological disorder and classify it automatically.

For this purpose, this research article analyses a collection of related works that are published for the purpose of pathological disorder detection. These approaches provide a baseline for the development of our technique. The in-depth analysis of the related words has provided valuable insight that has resulted in an effective mechanism for our approach. The design approach utilizes protocol Estimation and k-nearest neighbor along with artificial neural networks which are paired with two different classification techniques namely decision tree and random forest classification for the purpose of pathological disorder Classification. The approach will be effectively expanded and detail in the upcoming researches on this paradigm.

This research paper utilizes the section 2 for evaluation of the previous researches in the form of a literature survey. Section 3 elaborates about the proposed idea, whereas the section 4 discusses about the obtained results. Finally, section 5 provides the conclusion and offers the future direction for research.

## II LITERATURE SURVEY

To assess genetic syndromes, Amit David and Boaz Lerner [1] developed a support vector machine that categorizes actual cytogenetic signals from fluorescence

in-situ hybridization (FISH) photographs. The SVM structural threat minimization idea is used in this study to find the best structure for the classifier kernel and parameters. Outcomes show off the proper performance of the SVM in categorizing FISH signals in observation to other state-of-the-art machine learning classifiers, exhibiting the capabilities of an SVM-based genetic analysis framework.

For the prediction of cardiac illness, Yan Zhang et al. [2] used SVM with RBF kernel function, which is based on statistical learning theory. The Grid Search technique of improving criteria is used to choose the best kernel function parameters and fine important characteristics, resulting in the highest classification accuracy possible.

Fuzzy NN with SVM and ANN was used by Saeed Shariati et al. [3] to detect and diagnose hepatitis and thyroid diseases. Furthermore, they identified the variety and stage of disease, which includes 6 classes for hepatitis, namely Hep B (two stages), Hep C (two stages), hepatitis, and non-hepatitis, and 5 classes for thyroid disease, namely hypothyroid, thyrotoxicosis, Subclinical hypothyroid, Subclinical thyrotoxicosis, and non-presence of the thyroid. The good accuracies extend reach 98 percent for hepatitis illness and 99 percent for thyroid illness.

Hanaa Ismail Elshazly et al. [4] suggested a support vector machine classifier based on a genetic method for lymph disease analysis. The dimensions of the lymph diseases dataset contain 18 characteristics in the first stage, which are reduced to six aspects using GA. In the second step, a support vector machine with many kernel functions such as linear, quadratic, and Gaussian was utilized as a classifier. The effectiveness of the SVM classifier with each kernel function is evaluated using efficiency measures such as accuracy, sensitivity, specificity, AUC/ROC, Matthews Correlation Coefficient, and F-Measure.

D. Tomar et al. [5] used LST-SVM and Particle Swarm **Optimization**-PSO to create a successful Parkinson's disease analysis system. PSO is used to select characteristics and optimize parameters. In terms of accuracy, sensitivity, and specificity, the suggested process is compared to many other existing processes. The suggested Parkinson's disease evaluation approach outperforms existing approaches with 98 percent accuracy, according to empirical results.

SVM and Probabilistic NN were proposed by Fatemeh Saiti et al. [6] for the categorization of two thyroid illnesses from the thyroid disorder database: hypothyroidism and hyperthyroidism. To manage excessive and insignificant components, these algorithms rely on effective categorization algorithms regularly. The Genetic Algorithm was put to the test as a useful and robust framework for deciding on fair selections of factors that result in improved prognosis rates.

The affectivity of Bayesian classifiers in detecting the risk of cardiovascular disease was investigated by Alaa Elsayad et al. [7]. Two Bayesian network classifiers are implemented: Tree Augmented Nave Bayes and Markov Blanket Estimation, and their prognostic certainty serves as reference points for the Support Vector Machine. The exploratory results show that Bayesian networks with MBE have a 97 percent classification precision, whereas TAN and SVM units have 88 and 71 percent, respectively.

S Hongzong et al. [8] are interested in using SVM to classify coronary heart disease and non-coronary heart disease. SVM with a radial basis function kernel and linear discriminant evaluation are compared. Following that, the prediction precisions of SVM training and assessment units were 96 percent and 78 percent, respectively, and for LDA it was 90 and 72 percent. SVM and LDA had 92 and 85 percent cross-validated accuracy, respectively.

For the investigation of hepatitis disease, Javad Salimi Sartakhti et al [9] proposed a unique machine learning approach that incorporates support vector machines and simulated annealing. Using 10-fold cross-validation, the classification accuracy is determined. The suggested technique was found to have a classification precision of 96.25 percent.

D. Vassis et al. [10] presented a comprehensive review of the use of neural networks in automated medical prediction, with a focus on Support Vector Machines (SVMs), which are precise variants of neural functions. During the assessment, neural programs may be utilized to anticipate symptoms and diseases in many circumstances, whereas SVMs are gradually employed in clinical prognosis because to their unique categorization aspects..

## III PROPOSED METHODOLOGY

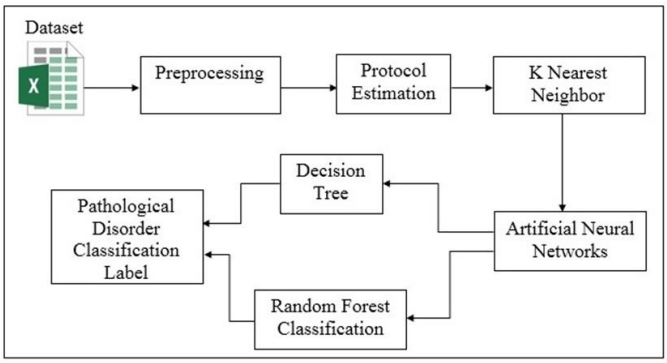


Fig 1: Proposed model for Driver distraction system

The proposed methodology for the Pathological Disorder classification has been depicted pictorially in the figure 1 given above. There are sequences of steps that have been implemented to achieve the methodology; these steps have been described below.

*Step 1: Data Collection and Preprocessing* – The initial step of the presented approach is targeted towards the collection of the data to provide it as an input to the system. The pathological disorders are very varied in nature which requires a number of parameters that need to be evaluated for accurate detections and classifications.

For the successful identification of the pathological disorder, a dataset containing the disorders and the SCADI dataset containing the data based on ICF-CY is being used. This dataset contains attributes collected for 70 children for self-care activities that are downloaded from

the URL - https://archive.ics.uci.edu/ml/datasets/SCADI. This dataset is interfaced with the java code through the use of the JXL library which converts the dataset into a double dimension list which can be easily utilized in the system.

The Graphical User Interface is designed to facilitate the registration of the user by taking basic information attributes such as name, Email address, Mobile number, username and Password. After successful registration, the user can log into the system using the username and password provided at the time of registration.

After successful authentication of the login credentials, the system provides the user to enter the parameters for the evaluation of the pathological disorders in the patient. These attributes include, Age, gender, washing oneself, caring for body parts, toileting, dressing, eating, drinking, and looking after one's health. The age is provided numerically whereas the gender is referred to as 1 for male and 0 for female. The other attributes accept binary input where 1 is a positive response and 0 is a negative response. These attributes are provided as an input to the system for further processing.

*Step 2: Preprocessing and Protocol Estimation:*The input dataset and the user input needs to be preprocessed before being provided to the system as an input. The dataset attributes are based on the International Classification of Functioning, Disability and health, which one of the classifications provided by the World Health Organization for the measurement of the disability and the health factors. These are especially useful in determination of the individual disability status.

The input dataset achieved from SCADI is derived from the ICF-CY, where the CY is the Children and Youth Version. This is specifically designed to diagnose and measure the functioning disability of children and young individuals. This resolution allows for a set of functions of the body and their respective impairments. The input dataset derives the qualifiers for self-care that is the most evident for pathological disorders in children and young people.

The dataset is cleaned in a way that the unnecessary qualifiers and the performance of these functions are eliminated. This leaves the only relevant data to be present for the purpose of our implementation protocols. This preprocessed dataset with the selected relevant attributes is provided to the subsequent step of the approach for further processing.

*Step 3: K Nearest Neighbor* – The preprocessed list attained in the previous step is utilized as an input in this step of the procedure. This step of the pathological disorder classification procedureperforms the K Nearest Neighbor clustering. This is performed toclassify the dataset given as an input into semantic groups to achieve accurate detection of the pathological disorders in an individual. This procedure is performed through the steps given below.

*Distance Evaluation* –The Euclidean Distance formula given in the equation 1 below is being used for the purpose of evaluation of the distance between the dataset attributes and the user input. The dataset is provided in the form of a list that is given as an input to this step of the procedure. The data stored in the rows of this list is used for the purpose of calculating the distance from the user input, and stored as the row distance at the end of the list. This is performed for all the entries in the list iteratively.

$$ED = \sqrt{\sum(ATi - ATj)^2} \_\_\_\_\_ (1)$$

Where,
ED=Euclidian Distance
$A_{Ti}$=Attribute at index i
$A_{Tj}$= Attribute at index j

*Centroid Estimation* –The list with the data and the respective row distances calculated in the previous step are used in this step as an input. The list is sorted in the ascending order of the Row distance and subjected to random data point selection. These selected data points are K in number and achieving the centroid. These data points are used to determine the row distance of the selected index. The obtained row distance is used further to achieve the boundary of the clusters.

The row distances achieved previously are used to form the average row distance of the entire list. These average row distances along with the extracted row distance of the centroid are used to determine the boundaries of the clusters in the cluster formation procedure given below.

*Cluster Formation* –The centroids and the average row distance acquired previously are provided as an input to this step for cluster formation. This step then calculates the boundaries of the clusters by addition and subtraction of the average row distance and the centroid row distance to attain the maximum and minimum values respectively. The data is then subjected to these boundaries to form the clusters which are aggregated into a cluster list. This cluster list sorted into a descending order and the top 3 clusters are provided to the next step as an input.

*Step 4: Artificial Neural Network* –The Artificial Neural Network is tasked with the process of identification of the pathological disorders. This step of the procedure provides the clusters obtained in the previous step as an input. These clusters are useful in determination of the pathological disorders based on the input data being provided by the user into the system. The cluster and the data are used to form the hidden layer and the output layer. The hidden layer formation is depicted mathematically in the equation 2 and 3 given below.

$$X = (AT1 * W1) + (AT2 * W2) + B1 \_\_\_\_\_ (2)$$

$$H_{LV} = \frac{1}{(1+\exp(-X))} \_\_\_\_\_ (3)$$

Where,
W1 ,W2 - Random weights,
( Other random weights set is also used like
{ W1,W2,W3,W4,W5,W6,W7,W8…,B1,B2… } )

AT1,AT2 - Attributes for which prediction probability is estimated
$H_{LV}$ – Hidden layer

The hidden layer probability is being calculated through the use of the equations given above. The random weights are being used for each of the attributes in a particular hidden layer and a bias weight for each one of them. These values along with the activation function called the sigmoid function provide the output layer values. These output layer values with the use of the target values allows for the calculation of the error probability values as an output. The equation 4 below demonstrates the calculation of the error probability values.

$$Error\ Probability = \sum \frac{1}{2}(T_0 - O_L)^2 \text{----- (4)}$$

Where,
T = Target Values
$O_L$ = Output Layer Values

The error probability rate is useful in correction of the values and improving the input attributes considerably. These values are collected together in the form of a list and provided for classification of the output.

*Step 5: Decision Tree & Random Forest Classification* – The Decision Tree and Random Forest Classification are being used to perform the classification. The inbuilt functions though the Support Vector Machine library in java is being used to implement the classification methodologies in an Ensembling format. The error corrected values are subjected to the classification of the pathological disorder labels provided by the Artificial Neural Networks. These classification methodologies provide the accurate output for the classification of the disorder and suggest the user areas which can be improved with the child.

## IV Results And Discussions

The presented approach for the purpose of pathological Disease classification through the use of Artificial Neural Networks along with Decision Tree and Random Forest Classification has been developed using Java programming language. The approach has been deployed using a laptop of the configuration consisting of an Intel Core i5 processor equipped with 6GB of RAM and 500 GB of Hard drive storage. The Database management and responsibilities have been handled by the MySQL Database server. The JXL API is being used to achieve workbook input into the java code.

The presence of any errors in the presented approach needs to be evaluated to understand any substantial reduction in the performance of the methodologies. These errors also allow for the effective realization of the implementation characteristic to reduce the incidences of faults or inconsistencies in the development of the pathological disorder classification methodology.

The prescribed approach provides an effective suggestion to the users of this approach for the pathological disorders and the sector that needs improvement. These suggestions need to be validated for their accuracy in the identification. This is a very complicated process that would require human intervention to validate the accuracy of the output. This is due to the fact the human has a better knowledge in understanding and analyzing the suggestion for the evaluation procedure.

Therefore, the Mean Reciprocal Rank is being utilized for the purpose of identifying the performance metrics of the proposed technique. The suggestion provided for the input attributes of the user need to be assessed for the relative accuracy that is achieved by the system in identifying the pathological disorders. This will be achieved by providing a rank for the respective suggestion provided by the user. This is due to the inherent ability for human to verify the solution objectively.

For the experimental evaluation 10 users are selected for the execution of the proposed methodology and the suggestions provided for the pathological disorder classification. The varying input is being provided for every run of the methodology and a rank between 0 and 6 is provided to the output suggestion that is provided.

The rank provided for the achieved suggestions is subjected to a reciprocal of the value. This is done to attain a value between 0 and 1. Where the value of 1 is the most accurate suggestion and the value of zero is the most inaccurate suggestion. These values are calculated for all of the users and a mean of the values is attained as depicted in the table 1 below.

| User | Pathological Disorder Classification Rank |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 0.5 |
| 4 | 1 |
| 5 | 1 |
| 6 | 0.5 |
| 7 | 1 |
| 8 | 0.5 |
| 9 | 0.5 |
| 10 | 1 |
| MRR | 0.8 |

Table 1: User Mean Reciprocal Rank (MRR) for blood Pathological disorder suggestion

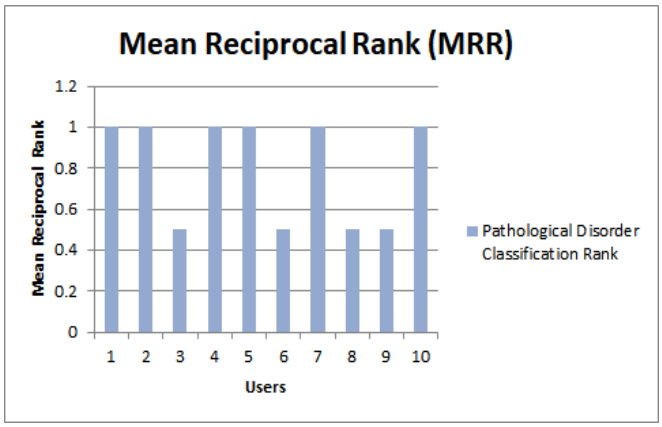


Fig 2: Graphical Representation of the MRR Values

The tabulated outcomes of Mean Reciprocal Rank are used to create a visual representation using a bar graph as given in the figure 2 given above. These values are useful in determining the accuracy of the approach which has been extremely satisfactory. These values have been depicted the accurate deployment of the Artificial Neural Networks along with the Decision Tree and Random Forest Classification. The output of these values has been considerable in the development of the approach and the classification of the pathological disorders. The methodology achieves an MRR of 0.80 which is extremely satisfactory.

## V. Conclusion And Future Scope

The paradigm of pathological disorder classification is one of the most essential needs of the hour. There has been an increase in the number of patients across the globe with pathological disorders. These disorders are highly difficult and complicated to predict and classify. This is due to the fact that these disorders are characterized with the varying symptoms that are difficult to categorize. Therefore, there is a need for the effective and useful technique for the automatic classification of the disorders. For this purpose the methodology proposed in this research article utilizes the user attributes and the dataset as an input which is provided for preprocessing. Once the data is preprocessed, it is effectively provided to the K Nearest Neighbors for the purpose of clustering. The clusters achieved for this purpose are given to the Artificial Neural Networks as an input. The ANN achieves error correction through the use of error probability estimation. The obtained values are then effectively classified using the Decision Tree and Random Forest Classification to achieve the suggestions as an output. The experimental outcomes through MRR describe the accuracy of the approach which is extremely satisfactory.

The pathological disorder classification approach prescribed in this research can be further augmented in the future research is through convergence of the methodology into an API for easier integration into existing systems.

## References


[1] Amit David and Boaz Lerner, “Pattern Classification Using A Support Vector Machine For Genetic Disease Diagnosis”, Proceedings. 2004 23rd IEEE Convention of Electrical and Electronics Engineers in Israel, pp 289~292.

[2] Yan Zhang, Fugui Liu, Zhigang Zhao, Dandan Li, Xiaoyan Zhou, Jingyuan Wang,”, Studies on the application of Support Vector Machine in diagnosing coronary heart disease”, Sixth International Conference on Electromagnetic Field Problems and Applications (ICEF), 2012, pp 1~4.

[3] Saeed Shariati, Mahdi Motavalli Haghighi,” Comparison of ANFIS Neural Network with Several Other ANNs and Support Vector Machine for Diagnosing Hepatitis and Thyroid Diseases”, International Conference on Computer Information Systems and Industrial Management Applications (CISIM), 2010, pp 596~599.

[4] Hanaa Ismail Elshazly, Abeer Mohamed Elkorany, Aboul Ella Hassanien,” Lymph diseases diagnosis approach based on support vector machines with different kernel functions”, 9th International Conference on Computer Engineering & Systems (ICCES), 2014, pp 198~203.

[5] D. Tomar; B. R. Prasad; S. Agarwal,” An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization” 9th International Conference on Industrial and Information Systems (ICIIS), 2014, pp 1~6.

[6] Bo Pang, David Zhang, Naimin Li, and Kuanquan Wang,” Computerized Tongue Diagnosis Based on Bayesian Networks”, IEEE Transactions on Biomedical Engineering, NO.10, VOL. 51, OCTOBER 2004, pp 1803-1810.

[7] Fatemeh Saiti, Afsaneh Alavi Naini, Mahdi Aliyari shoorehdeli, Mohammad Teshnehlab,” Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM”, IEEE 3rd International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2009, pp 1-4.

[8] Alaa Elsayad and Mahmoud Fakhr,” Diagnosis of Cardiovascular Diseases with Bayesian Classifiers”, Journal of Computer Sciences 2015, 11 (2): 274-282. DOI: 10.3844/jcssp.2015.274.282.

[9] Emre Omak, Ahmet Arslan, Ibrahim Turkoglu,” A decision support system based on support vector machines for diagnosis of the heart valve diseases”, Computers in Biology and Medicine 37 (2007)21 – 27.

[10] Javad Salimi Sartakhti, Mohammad Hossein Zangooei, Kourosh Mozafari,” Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM- SA)”, computer methods and programs in biomedicine 108 (2 0 1 2 ) 570–579.