IJIEMR Transactions, online available on 30th Jan 2024. Link :http://www.ijiemr.org/downloads.php?vol=Volume-13&issue=Issue01

10.48047/IJIEMR/V13/ISSUE 01/07

TITLE: **An Ensemble Machine Learning Model that Predicts Human Illnesses Based on Symptoms**

Volume 13, ISSUE 01, Pages:53-62

Paper Authors **Mr. S Koteswara Rao Yarlagadda1, Dr. S. Krishna Rao2**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

An Ensemble Machine Learning Model that Predicts Human Illnesses Based on Symptoms

Mr. S Koteswara Rao Yarlagadda[1], Dr. S. Krishna Rao[2]
[1]Assistant Professor, Department of Information Technology, Sir C R Reddy College of Engineering, ELURU
yskrao71@gmail.com
[2] Professor, Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, ELURU
skrao71@gmail.com

Abstract: Preventing and treating illnesses necessitates a prompt and accurate examination of all health-related issues. When diagnosing a serious illness, the standard procedure may not be adequate. The creation of a medical diagnosis system based on machine learning algorithms for disease prediction may enable more accurate diagnosis than the old technique. We built a disease prediction system with machine learning techniques. The processed dataset contains over 133*43 in testing and 133*4921 in training datasets. The diagnosis algorithm produces an output that depicts a person's potential condition based on their common cold, hepatitis, allergies, fungus, and symptoms. Developing strong classifiers by integrating all models produces improved results when compared to other approaches. The combination of the three models was projected to produce the same result. When a disease is diagnosed early enough, our diagnosis model may serve as a doctor, allowing treatment to begin on time and saving lives.

Keywords: K-Fold Cross validation, Confusion matrix, Support Vector Machine, Random forest, Gaussian Naïve Bayes Classifier.

Introduction

Medicine and healthcare are critical components of both the human condition and the economy. The world that existed just a few weeks ago has drastically changed from the one we live in today. Everything has become bizarre and graphic. The medical professionals are working as hard as they can to save people's lives in this virtual world, even if it means endangering their own. Certain rural populations lack access to healthcare facilities. Virtual doctors are board-certified physicians who prefer to practice virtually through phone and video consultations rather than in-person visits; nevertheless, they cannot treat patients in an emergency. Because they are incapable of making mistakes, machines are always regarded as superior to humans because they can complete tasks more quickly and accurately every time A disease predictor, often known as a virtual doctor, can diagnose any patient's condition without relying on human mistake. Furthermore, a disease predictor can be a godsend in situations like COVID-19 and EBOLA since it can diagnose a person's illness without any direct physical touch. There are several virtual doctor models available, but they lack the necessary level of accuracy because not all the necessary criteria are taken into account. The main objective was to create a variety of models and determine which one could provide the most accurate forecasts. Although machine learning algorithms differ in size and complexity, they all follow a similar fundamental framework. A number of machine learning-based rule-based strategies were utilized to recollect the creation and implementation of the predictive model. A number of machine learning (ML) methods were used to start several models, gathering raw data and splitting it into groups based on symptoms, age, and gender. After that, the data set was processed to create a strong model that combined ML models such as Random forest, SVM, and Gaussian Naive Bayes. Each model received the input parameters data-set during data processing, and the disease was returned as an output with varying degrees of accuracy. The model that achieved the highest accuracy level has been chosen.

LITERATURE REVIEW

In [1] Shraddha Subhash Shirsath, Prof. Shubhangi Patil has presented, “Disease prediction using Machine Learning over Big Data”. The concept's objective is to identify a region and gather hospital or medical data from that specific region. A machine learning algorithm is used in this process. The partial data is then obtained and reduced by identifying the missing data based on latent characteristics. The prior system victimized the CNN-MDRP at a level that was endlessly implemented after using the CNN-UDRP. The CNN-MDRP gets around CNN-UDRP's disadvantage. Both organized and unstructured hospital data are used by CNN-MDRP. When compared to earlier systems, the prediction made by the CNN-MDRP algorithm is more accurate. The concept's benefits include improved feature description and accuracy; however, this system's drawback is that this feature is limited to organized data, making it unsuitable for describing diseases.

In [2] Vinitha S, Sweetlin S, Vinusha H, Sajini S have put forth the idea of utilizing big data for machine learning-based illness prediction in order to get over the limitations of machine learning. The suggested idea is put to the test or tested using actual hospital data collections, such as daily updated data on hospital-oriented information, patient and doctor preferences for patient and doctor details, disease and disease-oriented data preferences, etc. The two primary issues with the current system that this technique addresses are (i) incomplete data and (ii) missing data. to reassemble the model of latent factors. The idea is to use the Machine Learning Decision Tree algorithm and the Map Reduce (MR) technique to obtain data from a hospital that gathered data from a forum known as "structured and unstructured data." Data partitioning is done using the MR method. It reports the likelihood of disease occurrences and achieves 94.8% with the conventional speed—quicker than CNN-UDRP.

In[3] Sayali Ambekar and Dr.Rashmi Phalnikar in may 2018 has presented the concept “Disease Prediction by using Machine Learning”. The idea of machine learning is used to retrieve information about diseases, and data analysis is used to accomplish the treatment procedures in these kinds of procedures. The decision tree is used extensively in disease outbreak prediction due to its high efficacy. This concept-based experiment demonstrates the relationship between the outcome and the symptoms of the condition, allowing a modified prediction model to be used to represent the data. If the idea selects the training set based on the symptoms of medical patients, it will use a decision tree, forecast, and then provide the patient's symptoms to obtain an accurate result for disease prediction. This idea is only used; that is, it only forecasts patient-related data quickly and affordably.

In[4] Lohith S Y, Dr. Mohamed Rafi. “prediction of disease using machine learning over big data”. able to create the foundation for a medical specialization This idea is used to create mass medical data—that is, data that has been expanded—from medical data. This concept's intended outcome is the storage of the most basic data inside the realm of medical huge data analysis, or "medical data analysis in massive collection." Compared to CNN-UDRP, it generates correctness and reaches 4.8% speed faster. Only these three types of data are covered: (a) structured data, (b) text data, and (c) combined structured and text data. The term "medical data oriented" is improved in this suggested system.

In[5] the author has presented, “personalized disease prediction care from harm using big data”, for healthcare analysis. Big data methods and the logic are used to analyze statistical analytics. The "disease recommendation system" that is being suggested is a technique that includes a specialized tool for constructing profiles. Certain information from the personalized individuals—doctor, patient, etc.—is required for the profile-making process. This idea is taken from and used in programmes such as Recommendation Engine and Collaborative Assessment (CARE). By analyzing the performance constraint, the CARE enhances the prediction of personalized diseases. The entire performance for patient-oriented big data is expressed by the CARE. It requires more time.

In [6] Gakwaya Nkundimana Joel, S. Manju Priya. “Use the Weighted Ensemble to Neural Network based Multimodal Disease Risk Prediction (WENN-MDRP) and have selection of Ant colony improved classifier for disease prediction over the large data concepts”. The Improved Ant Colony Optimisation (IACO) technique is used in the suggested concept to resolve the complexity of feature selection issues for huge data-based data analytics. The unheard technique, also known as the second WENN-MDRP technique, aids in the selection of the best features from medical data. When these two approaches are combined, there is a unique advantage of better prediction accuracy when compared to experimental methodologies. This idea only functions when there is sufficient time to complete the necessary tasks, such as (i) recall, (ii) accuracy, and (iii) precision. It chooses the best option that is conceivable without first investigating the potential.

In[7] 2018 Asadi Srinivasulu, S.Amrutha Valli, P.Hussain Khan, and P.Anitha introduced "Disease prediction in big data healthcare" using extended CNN. Predicting the risk of liver-oriented disease is the goal of the suggested system. Hence, the hospital dataset exclusively gathers structured data from liver illness information and is associated with diseases that are liver-oriented. The precision is obtained in the suggested system through the application of disease risk modelling. However, the risk prediction is more accurate when it takes into account the many aspects of medical data.

In[8] 2018 author has presented big data techniques in public health like, “Big data in public health: Terminology, Machine Learning, and Privacy”. Big data is utilized in medical field-oriented study, which also includes protection and hypothesis-generating research. This suggested system's interpretability does not use machine learning techniques.

In[9] Smriti Mukesh Singh, Dr. Dinesh B. Hanchate has presented the concept “Improving disease prediction by machine learning”, that is using machine learning and improving the disease prediction. This idea makes use of a genetic algorithm and recovers data—that is, missing data—from a dataset that also contains medical data. The two calculation terms used by this system are (i) KNN and (ii) SVM. The number of chronic illnesses rises. The medical data is used in the CNN-MDRP approach. The database contains personal and medical information as well as a thorough history of each patient. The logical data may be found with ease using RNN-based approaches. Both online and offline techniques are used in this system.

In[10] the author has presented the concept “Competitor Mining and Unstructured Dataset Handling Technique”, Competitive mining is discussed in this study along with similar publications. at last provided rival mining algorithms along with their benefits and cons. Comparing this paper's experimental outcome to others, C Miner++ produced the least amount of computing time.

METHODOLOGY

Based on an open-source dataset, we built an excel sheet with all of the symptoms for the linked illnesses. We had more than 133*43 in testing and 133*4921 in training datasets. Symptoms of a common cold, hepatitis, allergies, and fungal infection were fed into various machine learning algorithms.

Procedure:

Step 1: Determine whether the given dataset is balanced or not.

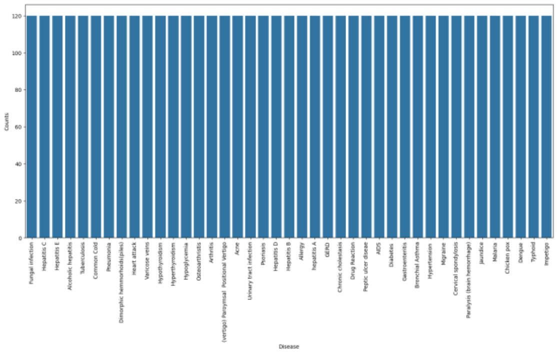

Figure 1: Checking if the given dataset is balanced or not.

Step 2: Using the Label Encoder, encode the target value into a numerical value.
Step 3: Splitting data for training and testing the model
Step 4: Defining the score measure using k-fold cross validation
Step 5: Create a robust classifier by merging all models with training and testing datasets for correctness.
Step 6: Fitting the model to the entire dataset and validating on the test dataset.
Step 7: Develop a symptom index dictionary to convert received symptoms into numerical values.
Step 8: Define the function based on the model's inputs and output.
Step 9: Calculate the final forecast by taking the average of all predictions.
Step 10: Assess the function.

K –Fold Cross Validation

When using K-Fold Cross Validation, the dataset is divided into k folds, or subsets. All of the folds are then used for training, with the exception of one (k-1) subset, which is used to assess the trained model. Using this approach, we iterate k times, reserving a distinct subset for testing each time.

It producing cross validation score for the models:

```
========================================
======================
SVM
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
========================================
======================
Gaussian NB
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
========================================
======================
Random Forest
Scores: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Mean Score: 1.0
```

Confusion Matrix

The confusion matrix is a matrix used to evaluate the performance of classification models on a given set of test data. It can only be calculated if the true values of the test data are known.

Table 1: Performance metrics

N=total Predictions	**Actual : Positive**	**Actual : Negative**
Predicted : Positive	True Positive	False Positive
Predicted : Negative	False Negative	True Negative

Classification Accuracy

It is an important parameter for determining the accuracy of classification tasks. It indicates how frequently the model predicts the proper output. It can be computed by dividing the number of right predictions made by the classifier by the total number of predictions made by the classifiers.

Accuracy = TP+TN / TP+FP+FN+TN

Precision

It can be defined as the number of correct outputs produced by the model or the proportion of positive classes predicted correctly by the model that were actually true.

Precision = TP / TP+FP

Recall (or) Sensitivity

It is defined as the proportion of total positive classes that our model properly predicted. The recall should be as high as possible.

Recall = TP / TP+FN

F-Score

Comparing two models with low precision and good recall, or vice versa, is difficult. So, for this reason, we can use the F-score. This score allows us to examine recall and precision at the same time. The F-score is greatest when the recall equals the precision.

F-measure=2*Recall*Precision /Recall + Precision

Gaussian Naïve Bayes Classifier

Gaussian Naïve Bayes assumes that continuous numerical attributes follow a regular distribution. The property is initially segmented based on the output class, and the variance and mean of each class are then calculated.

Bayes' Theorem

Bayes' theorem, often known as Bayes' Rule or Bayes' Law, is used to calculate the probability of a hypothesis based on prior knowledge. It all depends on the conditional probability. The formula for Bayes' theorem is provided below:

$$\mathrm{P(A|B)}=\frac{P(B|A)P(A)}{P(B)}$$

The probability of A|B is Posterior probability: The likelihood of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: P(A) represents the likelihood of a hypothesis being true based on available evidence. Prior probability is the probability of a hypothesis before witnessing the evidence.

P(B) is Marginal probability: probability of evidence

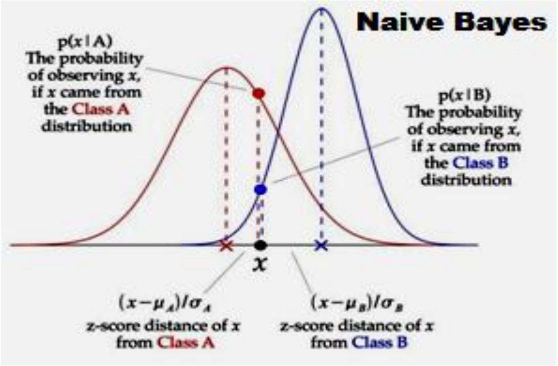
Naive Bayes
p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Figure 2: Naïve Bayes Classifier

This is the probability that feature x will occur; it transforms a probability function into a distributed function with a total probability of one.

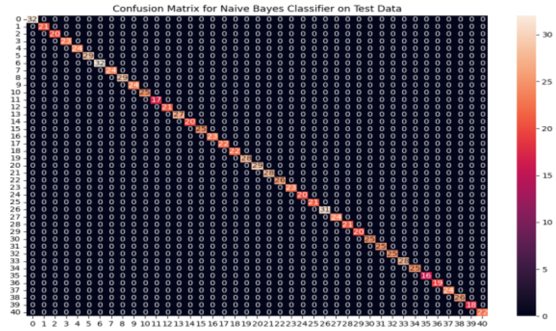

Figure 3: Confusion matrix for Naïve Bayes classifier using test data.

Accuracy on train data by Naive Bayes Classifier: **100.0**
Accuracy on test data by Naive Bayes Classifier**: 100.0**

Random Forest Classifier

The mean squared error (MSE) formula is used by the Random Forest algorithm in machine learning to tackle regression problems. The MSE formula determines each node's distance from the estimated actual value. This aids in selecting the branch that is best for the forest.

$$MSE = \frac{1}{N}\sum_{(x,y)\in D}(y - prediction(x))^2$$

Step 1: From a given training set or data, choose random samples.
Step 2: For each training set of data, this algorithm will build a decision tree.
Step 3: The decision tree will be averaged to determine the winner.
Step 4: Choose the predicted result that received the most votes to be the final outcome.

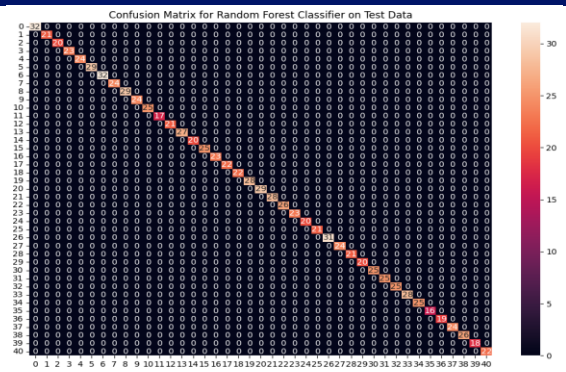

Figure 4: Confusion matrix for Random Forest classifier using test data.

Accuracy on train data by Random Forest Classifier: **100.0**
Accuracy on test data by Random Forest Classifier: **100.0**

Support Vector Machines Classifier

The SVM algorithm's primary goal is to locate the best hyper plane in an N-dimensional space that may be used to divide data points into various feature space classes. The hyper plane attempts to maintain the largest possible buffer between the nearest points of various classes. The number of features determines the hyper plane's dimension. The hyper plane is essentially a line if there are just two input features. The hyper plane transforms into a 2-D plane if there are three input features.

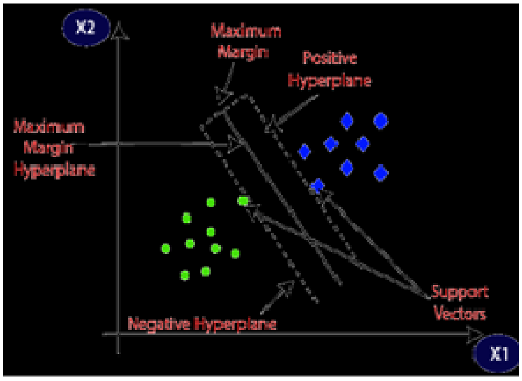

Figure 5: SVM Classifier

SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM algorithm can be used for Face detection, image classification, text categorization, etc. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane

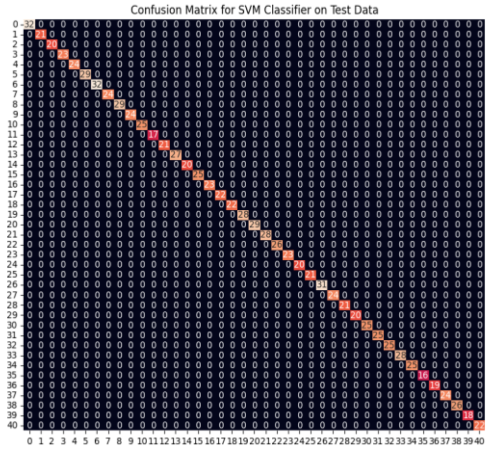

Figure 6: Confusion matrix for the Support Vector Machine classifier on test data.

Accuracy on train data by SVM Classifier: **100.0**
Accuracy on test data by SVM Classifier: **100.0**

SYSTEM ARCHITECTURE

In the proposed approach, the data is separated into train and validation sets. The train data is preprocessed before being split into 80% train data and 20% test data. Then, using these train and test data, k-fold cross validation is performed for model selection. Three distinct machine learning methods, such as support vector machines, Random Forest, and Gaussian Nave Bayes classifiers, are employed to forecast the final prediction.

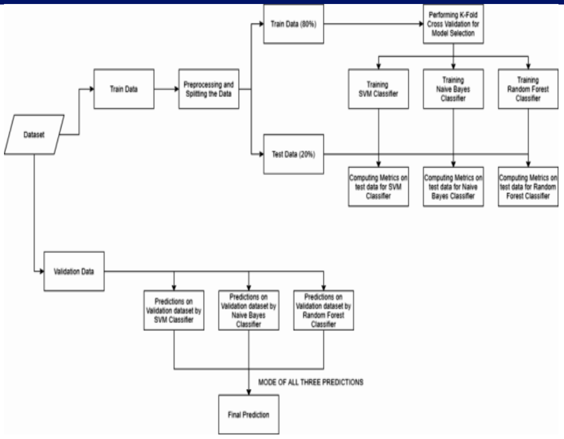
Dataset
Train Data
Preprocessing and Splitting the Data
Train Data (80%)
Test Data (20%)
Performing K-Fold Cross Validation for Model Selection
Training SVM Classifier
Training Naive Bayes Classifier
Training Random Forest Classifier
Computing Metrics on test data for SVM Classifier
Computing Metrics on test data for Naive Bayes Classifier
Computing Metrics on test data for Random Forest Classifier
Validation Data
Predictions on Validation dataset by SVM Classifier
Predictions on Validation dataset by Naive Bayes Classifier
Predictions on Validation dataset by Random Forest Classifier
MODE OF ALL THREE PREDICTIONS
Final Prediction

Figure 7: Machine-learning-based disease prediction

RESULT & DISCUSSION

Based on the preceding predictions, these machine learning models anticipate the same thing as the combined three machine learning models with perfect accuracy.

Table 2: Individual final predictions for several models

Models	SVM	RF	GNB
Result	Fungal infection	Fungal infection	Fungal infection

Final Prediction:
{'rf_model_prediction': 'Fungal infection', 'naive_bayes_prediction': 'Fungal infection', 'svm_model_prediction': 'Fungal infection', 'final_prediction': "Fungal infection"}

Table 3: Final projections of the combined three models.

Models	**SVM + RF+ GNB**
Result	Fungal infection

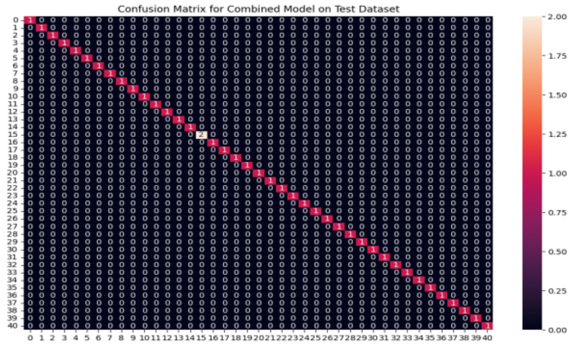

Figure 8: Confusion matrix for mixed model for the test dataset.

Accuracy on Test dataset by the combined model: **100.0**

CONCLUSION

The study proposed a strategy for forecasting sickness based on a patient's common cold, allergies, fungal infection, hepatitis, and symptoms. When the three models worked together to forecast diseases using the previously specified criteria, and predicted the same disease in all three models, the best accuracy of 100.0% was achieved. Almost all machine learning models generated accurate findings. As soon as we predict the sickness, we can easily assign the pharmaceutical resources required for therapy. This strategy would speed up the healing process and help to reduce the costs of treating the sickness. The goal of this study is to predict disease using symptoms. This project configures the device to receive the user's symptoms as input and output a disease prediction.

REFERENCES

[1] Prof. Shubhangi Patil, Shraddha Subhash Shirsath, "Disease Prediction Using Machine Learn.Over Big Data." Innovative Research in Science, Engineering, and Technology: An International Journal, [2018]. Online ISSN: 2319-8753; print ISSN: 2347-6710.

[2] Vinitha S, Sweetlin S, Vinusha H, and Sajini S. "Machine Learning Over Big Data for Disease Prediction." An International Journal of Computer Science & Engineering (CSEIJ), Vol.8, No.1, [2018], doi:10.5121/cseij.8101.

[3]Sayali Ambekar and Dr. Rashmi Phalnikar. "Machine Learning-Based Disease Prediction." May 18, special issue of International Journal of Computer Engineering and Applications, Volume XII. ISSN: 2321-3469.

[4] "Prediction of Disease Using Learning over Big Data - Survey," Lohith S. Y., Dr. Mohamed Rafi. International Journal of Communication Engineering and Computer Science: Future Revolution. 2454-4248 ISSN.

[5] "Personalised Disease Prediction Care from Harm using Big Data Analytics in Healthcare," by J. Senthil Kumar and S. Appavu. DOI:10.17485/ijst/2016/v9i8/87846, Indian Journal of Science and Technology, vol. 9(8), [2016]. Print ISSN: 0974-6846; Online ISSN: 0974-5645.

[6] Nkundimana Gakwaya Joel, S. Manju Priya. "Weighted Ensemble to Neural Network Based Multimodal Disease Risk Prediction (WENN-MDRP) Classifier for Disease Prediction Over Big Data with Enhanced Ant Colony on Feature Selection." Engineering & Technology International, 7(3.27) (2018) 56-61.

[7] The authors of "A Survey on Disease Prediction in Big Data Healthcare using an Extended Convolutional Neural Network" are Asadi Srinivasulu, S. Amrutha Valli, P. Hussain Khan, and P. Anitha. [2018] National conference on emerging trends in engineering sciences, management, and information.

[8] Mooney Stephen J. and Pejaver Vikas. The 2018 Annual Review of Public Health article is titled "Big data in public health: Terminology, Machine Learning, and Privacy."

[9] Dr Dinesh B. Hanchate and Smriti Mukesh Singh. "Improving Disease Prediction by Machine Learning." 2295-0056 for e-ISSN, 2395-0072 for p-ISSN.

[10] Joseph, Nisha, and B. Senthil Kumar. “Competitor Mining and Unstructured Dataset Handling Technique” by Machine Learning.

[11] A Survey on Disease Prediction by Machine Learning over Big Data from Healthcare Communities International organization of Scientific Research 59 | Page

[12] Journal of Network Communications and Emerging Technologies (JNCET) www.jncet. org 8.2 (2018). “Investigation Using Data Mining Technique”.

[13] B. Senthil, Kumar. "Adaptive Personalised Clinical Decision Support System Using Effective Data Mining Algorithms." https://www.jncet.org/8.1 (2018) Journal of Network Communications and Emerging Technologies (JNCET).

[14] B. Senthil Kumar, Asha, and Unnikrishnan. "Biosearch: A Domain Specific Energy Efficient Query Processing and Search Optimisation in Healthcare Search Engine." https://www.jncet.org/ 8.1 (2017) Journal of Network Communications and Emerging Technologies (JNCET).

[15] B. Senthil, Kumar. "Adaptive Personalised Clinical Decision Support System Using Effective Data Mining Algorithms." https://www.jncet.org/ 8.1 (2017) Journal of Network Communications and Emerging Technologies (JNCET).

[16] Senthil Kumar, B. "Data Mining Methods and Techniques for Clinical Decision Support Systems." Journal of Network Communications and Emerging Technologies (JNCET) 7.8 (2017) www.jncet.org.

[17] "Identification of Diabetes Risk Using Machine Learning Approaches," Sreejith, B. Senthil. www.jncet.org 7.8 (2017) Journal of Network Communications and Emerging Technologies (JNCET).

[18] Varma Bhavitha and Senthil B. "A Different Type of Feature Selection Methods for Text Categorization on Imbalanced Data." www.jncet.org 8.1 (2017) Journal of Network Communications and Emerging Technologies (JNCET)

AUTHOR PROFILE

Mr. S KOTESWARA RAO YARLAGADDA, A well-known teacher with an M.Tech (CSE) from JNTU Kakinada University works as an Assistant Professor in the Department of IT at Sir C R Reddy College of Engineering. He is an active member of ISTE. He has 14 years of teaching experience at several engineering universities. He has a number of publications to his credit, including national and international conferences/journals. His research interests include machine learning, deep learning, information security, cryptography, network security, and other breakthroughs in computer applications.

Dr. S. Krishna Rao is a well-known instructor who earned a PhD in Computer Science and Engineering from ANDHRA University. He is a professor in the Computer Science and Engineering department of Sir C. R. Reddy College of Engineering. He is an active life member of ISTE, the Institute of Engineers, and CSI. He has supervised ten researchers throughout the course of his 22 years of teaching and 10 years of research. He is an active member of the BOS and Editorial boards at several engineering universities. He has contributed to several national and international magazines and conferences. His research interests include network security, information security, mobile cryptography, wireless sensor networks, machine learning, deep learning, data structures, and other computer-related improvements.