International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Diagnosis of Vector Borne Disease using Various Machine Learning Techniques

Salim G. Shaikh[1] , Dr. B. Suresh Kumar[2], Dr. Geetika Narang[3] Prof. N.N.Pachpor[4]

Submitted: 16/11/2022 **Accepted:** 18/02/2023

Abstract: Vector-borne diseases (VBDs) are one of the most serious human health issues, impacting millions of people each year in every corner of the globe. Multiple decision-making techniques are employed in this study to give a better diagnosis of VBDs. It assesses alternative illnesses with opposing symptoms. It is difficult to precisely define the weight of criteria and the ranking of alternatives (diseases) for each criterion. The proposed method is used to diagnose VBDs such as malaria, chikungunya, and dengue fever. In this paper, we proposed a prediction of VBD using various supervised machine learning classification algorithms. The Weka 3.7 machine learning framework has been used for the classification of data. The algorithms used, such as SVM, Naive Bayes, Adaboost, decision tree, ANN, etc., In extensive experimental analysis, we observed the SVM prediction had better detection and classification accuracy over the other machine-earning classes. For evaluation, we used 3000 records of patient data. The modified SVM (mSVM) achieves 100% accuracy for different cross validations.

Keywords: Medicine diagnosis, supervised machine learning, detection and classification, disease detection, vector borne disease.

1. INTRODUCTION

The majority of medical studies are focused on predicting and identifying disease-causing variables. The health researchers make suggestion proof for therapy based on good estimates. Numerous diagnostic approaches and decision support tools have been presented in recent years to improve physicians' skills and knowledge in accurately identifying and recognizing disorders. According to new study, Artificial Neural Networks (ANNs) are widely used in the field of healthcare data mining. For the advantage of illness diagnostic techniques, classical pattern matching approaches are substituted by Multi - layered NN (Neural Networks), which are regarded an important device for Multi layered Neural Network learning. Individuals presently suffer from a variety of illnesses of the climate and its lifestyle choices. As a result, predicting sickness at a preliminary phase gets crucial. The much more difficult duty is to predetermine and diagnose illness. The goal of this study is to anticipate vector-borne infection and develop a recommender. Recommendation model, in practice, are a collection of methods and procedures that can propose "appropriate" content to consumers. The recommended items should generally be as relatable as practicable, so that they can be engaged with: Videos online, media articles, web items, and so forth. The participants were presented with more relevant instances, which are ordered as per their relevance. The recommendation systems must assess relevance, which is primarily based on past information. If you've did watch bull clips on YouTube, Internet will begin to display you more bull videos featuring similar ideas and topics. Collaborative filtering as well as content-based methods and hybrid technique are basic categories of recommendation system. Collaborative Filtering Technique is further divided into Model based technique and Memory based technique which is illustrated in the following figure. Consider the following figure 1 demonstrates a recommender Model.

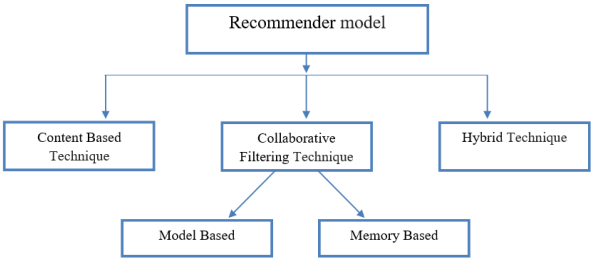

Figure1: Recommender Model in machine learning

[1]*Research Scholar, department of CSE, Amity University, Jaiapur, India.*
[1]*Assist. Professor, Computer Department Kalsekar Technical Campus, New Panvel, Mumbai.*
[2]*Associate professor department of CSE Sanjay Ghodawat University Kolhapur.*
[3]*Associate professor Head of Dept CSE. TCOER,Pune, India.*
[4]*Assist. professor IIMS, Pune, India.*
[1]*Shaikhsg2@gmail.com ,* [2] *sureshkumarbillakurthi@gmail.com ,*
[3]*geetikanarang2020@gmail.com,* [4]*nnpachpor@gmail.com*

For recommendation model, collaborative filtering approaches are procedures that are exclusively based on previous user engagement and the targeted objects. As a result, all previous information about user engagement with objects in the scene will be fed into a collaborative filter system. This information is usually recorded in a matrix, with rows representing users and columns representing objects. Memory and model based techniques are the two most popular forms of collaborative filtering systems.

Memory-based Methods: This method are also known as neighborhood collaborative filtering. Scores of user-item pairs are basically estimated based on their neighborhoods. Consumer as well as object-based collaborative filtering is two types of collaborative filtering. Consumer -based simply implies that similar consumers will produce similar and powerful suggestions. Object-based collaborative filtering suggests items based on their similarity, which is estimated depending on user rankings.

Model-based Techniques: It is a machine-learning-based predictive models. The model's parameters are specified using set of features in resolving an optimization issue. DT (Decision trees), rule-based techniques, latent factor frameworks, and other prototype techniques are examples of model-based techniques.

The key benefit of adopting collaborative filtering models is their ease of implementation and good standard of coverage. It had the advantage of capturing complex traits (this is certainly relevant for latent factor systems) while requiring no knowledge of the item content. Since there seems to be no contact between the consumer and the object, the fundamental downside of this approach is that it is not suitable for proposing new things. This is known as the "cold begin" issue. On extremely sparse information, memory-based techniques are known to underperform. Various collaborative filtering algorithms include: YouTube content recommendations to users — proposing films to you depending on what other viewers have subscribed to and viewed. Course Era course recommendations are depending on the outcomes of other people who have completed the same programs you have.

Suggestions are generated by content-based systems depending on the customer's wishes and biography. They strive to connect users with goods they've previously liked. The degree of correlation between products is usually determined by the features of items that the user likes. Unlike many of the collaborative filtering methods, which rely on rankings between the consumer as well as other individuals, content-based methods concentrates on the targeted user's own score. In general, the content-based method creates suggestions by combining several datasets.

The corresponding data sources are required for the basic forms of content-based solutions (these needs can grow depending on the severity of the solution you're attempting to develop):

Item-level information source — we have to have a reliable and valid data about the object's properties. The more information you have about the object, the better for your strategy it will be.

User-level data source – we need feedback from users on the object for which you're making suggestions. This type of evaluation could be subtle or clear. The more customer input you can collect, the better your method performs.

When there aren’t enough review data available, content-based systems are the best option for suggesting things. This is because the consumer may have rated other things with identical features. As a result, even if it's not a lot of information, a model is expected to provide suggestions using the evaluations and item features. The drawbacks of content-based systems are twofold. Depending on the things / information the user has eaten, the suggestion provided is "clear." This is a negative since the user will never again be suggested an item if they have never dealt with it. If you've never seen a thriller novel, for instance, you'd never suggest one using this method. This is due to the fact that the paradigm is consumer specific and does not take advantage of the information from other customers. This lowers the variety of suggestions, which is a bad thing for so many firms. They're useless when it comes to making suggestions to new subscribers. A record of explicit or implicit consumer level information for the objects is required while developing a project. To construct accurate findings without overloading, you should have a training set of evaluations accessible.

Different methodologies of recommender system get their own strengths and weaknesses in the Hybrid Recommender model. Whenever employed in isolation, several of these strategies can appear limiting, particularly whenever multiple data sources are accessible for the issue. Hybrid recommendation system is those that make use of a variety of information sources to make provision of adequate. Hybrid solution integrates several paradigms to overcome the shortcomings of one approach. Ultimately, this mitigates the drawbacks of utilizing individual methods and facilitates in the generation of more reliable suggestions. Users will receive better robust and tailored suggestions as a result of this. Such algorithms are typically computationally demanding, and they necessitate a large database of evaluations as well as other characteristics to maintain. It's tough to train and deliver specific proposals with revised goods and rankings from diverse users without up-to-date statistics (user interaction, reviews, etc.).

In numerous vector-borne infections like dengue virus, urgent diagnosis is critical. Dengue and severe dengue have no medical intervention. Dengue has become much more common over the world in recent years. 50 percent of the worldwide people are currently endangered. Each and every year, between 100 and 400 million illnesses are anticipated. Internal bleeding as well as organ destruction are possible side effects of serious dengue infection. Women who contract dengue while pregnant may pass the infection on to their babies during delivery. Because of the complex interaction among environmental related variables, analyzing and predicting Vector-Borne Infections appears to be a difficult challenge. The major goal of this study is to design an efficient forecast and recommender prototype for Vector-Borne illness utilizing ML (machine learning) techniques. Using only a variety of machine learning (ML) methods extract the features and problems related with Vector-Borne illness prediction systems that can be useful for creating a recommender in the health sector.

To determine the best feature set for better Vector-Borne illness prediction. To develop either forecasting as well as recommender system that can be used to assist humanity. Evaluate the suggested and traditional models' effectiveness using a variety of assessment indicators.

Multiple internet health forums on medical problems nowadays are accessible in the information technology field (In 2019, Vijayakumar. V et. al.) [1]. Clinical diagnosis is essential in the initial stages of diseases for example and Cancer and other disorders. Medical services are being established to meet the demands of a growing populace. Physicians can use the resources provided by the medical recommendation systems to diagnose diseases. HRS also advises patients on illness identification and then how to stay healthy. The purpose of this study is to use machine learning (ML) to design an efficient health recommendation system for diagnostic testing. Individuals from various demographic groups are affected by various forms of deadly diseases. Dengue fever is among the diseases. Dengue is still a life-threatening virus which has traveled to numerous parts of the world (In 2019, Inokuchi M. et. al.) [2]. Dengue fever is an acute disease caused by the reproduction of Aedes mosquitos. It's known as breakbone flu, and it's spread by the Aedes aegypti mosquito. Allergic reactions, muscular joint soreness, headaches, minty flavor, decreased platelet levels, and hypotension are the main signs of this condition (In 2017, Iqbal Naiyar et al.) [3]. There are no medicines or specific drugs available for this therapy. It occurs in a cyclic pattern and lasts for a total of 2 weeks in the body. (In 2019, Taneja P. et. al.) [4]. It is caused by a virus found in the spit of the Aedes mosquito, and it is transmitted to fit people through bite, where it mixes with bodily excretions. As a result, the dengue viral cycle begins, with replication taking place within WBC (white blood cells). If a serious infection occurs, the bone marrow as well as liver are damaged, and BP and artery flow are reduced (In 2003,Guzman. M. et. al., 2003) [5]. The platelet count drops as a result of the bone marrow disease, as well as serious hemorrhage occurs.

Because of the intricate interplay between environmental related variables (In 2010,San. J. L et al.) [6], identifying and predicting aedes illnesses appears to be a hard process (In 2014, Shepard. D. S. et. al.) [7]. As a result, the occurrence trends in different locations alter. The data analysis is complicated by the paucity of environmental parameters and lengthy background data. Variations in the immunological population distribution caused by the cycling of many virus types, as well as variety of human demographic influences such as brief movement, migration, and fertility rates, define the environmental fluctuations (In 2005, Ibrahim M et. al.) [8](In 2010, A. L. V. Gomes et al.) [9]. Rapid recognition of dengue sickness is critical in the global public healthcare system, and ML (machine learning) technologies assist physicians in identifying and predicting illnesses at an initial phases (Guo P. et. al., 2017) [10]. With the exception of improving classification performance, it also reduces diagnostic test pain, expense, and diagnosis time (In 2018, Carvajal T. M. et. al.) [11].

The Research paper is divided into Section as follows. Section II describes the related work done by the various researchers related to the identification of the vector born disease using Machine Learning Techniques. Section III focuses on Proposed Methodology of the developed model. Section IV concentrates on algorithm design of the presented research work. Ultimately Section V and Section VI show the experimental test outcomes and further conclude the proposed work respectively.

2. LITERATURE SURVEY

The majority of medical studies are focused on predicting and identifying disease-causing variables. Health researchers present advice proof for therapies depends on the forecast (In 2015, Isinkaye. F.O. et al.) [12]. Numerous detection systems including decision support tools have been presented in recent years to improve doctors' knowledge and skills in accurately identifying and identifying illnesses. According to new analysis, ANN (Artificial Neural Networks) are extensively employed of health information mining, and many methods are created utilizing the Artificial Neural Network due to its adjustment, parallel operation as well as forecasting capabilities. For the advantage of illness diagnostic procedures, classical pattern identification approaches are substituted by Multilayer NN (Neural Networks), which are regarded a useful weapon for MLNNs development. The items and patients are the 2 most important entities in

recommenders. Participants expressed recommendations for certain things, which are discovered using the collected data. The information gathered is used to create an utilitarian matrix that shows the importance of each patient-item combination and indicates the person's favorability for those objects. As a result, recommendation system algorithms are separated into two types: object as well as patient-oriented. Patients provide evaluations and object selections in the environment of a patient-oriented recommender (Martinez, B et al., 2015) [33]. The object can be recommended to the client utilizing a client recommendation system engine that takes into account the sufferers' similarity. In the instance of an object oriented recommender system, resemblance between objects is employed to make clinical forecasts. The primary prediction task, in the eyes of recommendation model, is information gathering.

In 2021, K. Indhumathi et al. [13] published a survey on seasonal prediction and diagnosis associated with changing climate employing big data, in which the researcher explained the Gaussian Process Regression Conceptual Model. The paper's key result is that climate change plays an essential part in bacterial contamination. Seasonal infections such as dengue fever, diarrhea, salmonella, and giardia lamblia have grown as a result of the excessive heat. The problem with the Gaussian Process Regression Method is that it is only useful for big datasets. ML (Machine learning) algorithms are being used to forecast dengue outbreaks in Selangor, Malaysia. In the forecasting of dengue infection associated with climate data, the researcher examined the methodologies CART, Artificial Neural Network (ANN), SVM (Support Vector Machine), and the NB (Naive Bayes) method. Climate features like temperature, wind velocity, moisture, and wetness were employed within every system as the report's outcome. According to the findings, the Support Vector Machine (linear kernel) had the highest forecasting accuracy (Accuracy is 70 percent, Sensitivity is around 14 percent, Specificity is about 95 percent, Quality is up to 56 percent) It is possible. To construct a dengue forecasting models, study intends to investigate enhancing or applying environment inspired methods. In 2021, Rayner Alfred et. Al [14] conducted research on the propagation of pathogens and machine learning (ML) algorithms for the fatal virus. In this paper, the article mentioned alternate methodologies including the Modified Apriori Technique for detecting dengue illness and Forecasting dengue epidemic Hybrid Neural Network (HNN), Artificial Neural Network, as well as Non-Linear Recurrence for forecasting dengue epidemic (NLR).The kind of data and characteristics employed are also recognized as a consequence of this paper, Identification on Dengue Infections using Modified Apriori Technique (ACC is 0.750), as well as Meteorology and Epidemiological information are determined to be among the most valuable databases for forecasting and identifying epidemics. There is always room for more research into the capabilities of prediction or hybrid methods depending on DL (Deep Learning) techniques that use multi-source information, as these will be demonstrated to boost the effectiveness of the underlying system. The research of machine learning (ML) methods for disease diagnosis was conducted by N. Reddy et al. [15] in 2021 for illness identification, Linear Regression (LR), Support Vector Machine, KNN, RF (Random Forest), DT ,Nave Bayes were examined. For the analyzing of liver illness, methods such as Back Propagation, SVM, RF ,NB, KNNNaive Bayes are utilized; amongst which, Naive Bayes had the higher precision of 95.1 percent. The study examines the strategy in relation to chronic diseases such as liver and kidney illness. The similar method can be used to verify the effectiveness of seasonal illness.

In the year 2020, alias Balamurugan et al. [16] enhanced classification efficiency as well as accuracy rate for patient monitoring technologies. Entropy Weighted Score based Optimal Ranking (EWSBOR) method was the name of the novel feature selection technique. It was quite helpful in terms of forecasting and analyzing health information. The sickness was primarily caused by the optimal extracted features. The dengue information was analyzed from various territories in Tamil Nadu's Thanjavur district. The results outperformed standard procedures. Using the help of the dengue database in Colombia, Ye J. et al. [17] distinguished many spatial conditionally autoregressive techniques in 2020. The research methodology included a new data modification. It combined Bayesian spatio-temporal modeling approaches with binary and Poisson-log normal techniques. The proposed framework was capable of capturing the information's modifications. The nighttime land surface temps, daylight, altitude, and density of population were all considered relevant factors. The potential dengue epidemic zones have been determined. The plant and moisture factors have no bearing on the strategy adopted. It demonstrated a spatial likely to affect linked to the slope in altitude. These findings could be used in subsequent dengue studies conducted.

Mussumeci E. et al. [18] established machine learning (ML) algorithms including Random Forest regression (RFR) and LASSO for integrating variable selection in 2020. The Long Short Term Memory used 790 cities to forecast the weekly incidence of dengue fever in Brazil. Multimodal sequence was employed as predictors, as well as sequence from the same locations have been used to capture the geographic transmission of infection element. In the prospective frequency forecast of dengue in regions of various shapes and sizes, this technique performed welIn 2019, Chakraborty T. et al. [19] present a hybrid

approach for detecting nonlinearity as well as linearity in data that combined the Neural Network Auto Regressive and Auto Regressive. The ARIMA technique was used to eliminate the linear proclivities in the information, and the depreciable amount were transferred to the Neural Network Auto Regressive strategy. This strategy was tested with three dengue period large datasets, so it outperformed conventional methods in order to achieve accurateness. With the use of a recently designed hybrid technique, dengue infections were accurately forecasted over duration. Appice A. et al. [20] developed a holistic ML (machine learning) technique for examining the temporal aspects of dengue information and temp in 2020, and used this information to generate dengue forecasts ambient temperature. A unique multi-stage combination comprising trend oriented temporal grouping, auto encoding as well as window oriented information representation was used to derive temporal aspects from past records. A pattern association oriented KNN ne classifier was used to make the forecast. Dengue hemorrhagic sickness occurrences were collected in 32 regions across Mexico between 1985 and 2010. It was first proved in both forecasting as well as simple regression. Gambhir S. et al. [21] developed a screening approach for detecting dengue fever sooner. A PSO-Artificial Neural Network oriented diagnosing approach was presented, with the PSO optimizing the ANN method's bias or load. This method was used to identify dengue cases. AUC metrics, failure rate, selectivity, sensitivities, and correctness were used to determine effectiveness. PSO, Naive Bayes, Decision Tree, and Artificial Neural Network were used to evaluate the created technique to numerous previous approaches. This strategy proved effective for predicting dengue fever in the early stages.

In 2019, Mello Román et al. [22] distinguished between 2 ML (Machine Learning) algorithms for diagnosing diseases: SVM (Support Vector Machines) and ANN. The data though was an actual database relating to patients diagnosed collected from the Paraguayan public healthcare system between 2012 and 2016. With less modifications in 30 different divisions of the data, the Artificial Neural Network obtained 97 percent specific, 96 percent sensitive, and 96 percent reliability. For specificity, sensitivity, and accuracy, Support Vector Machine (SVM) achieved results of over 90 percent. With the assistance of the RBM (Restricted Boltzmann Machine)- CNN (Convolutional Neural Network) DL approach, a smart HRS has indeed been suggested. It suggested using big data analytics for a more improved healthcare recommendation systems algorithm deployment and identified a possible chance for the medical sector. In the framework of a tele-health system, a current situation was turned into a tailored approaching respect of MAE (Mean Absolute Error) as well as Root Square Mean Error, the presented scheme produced less mistakes than existing strategies. In 2018, Ivens Portugal et al. [23] identified trends with the use or development of machine learning (ML) techniques in recommendation system, as well as unresolved problems in the use or study of machine learning (ML) method. The author mention many aspects of a collaborating method, material filtration strategy, and a hybrid approaches technique, which discusses how and where to propose objects depending on the customer and object data. More research towards the use of Cluster, Support Vector Machine (SVM), and Ensemble in RSs can indeed be explored in future to see what effects their usage, effectiveness, and usefulness have.

In 2019, Vaishali S. Vairale et al. [24] present some current findings in the realm of nutrition and lifestyle that concentrates on tailored suggestions solely on medical data, bearing in mind their selection of diet and physical activity, as well as dietary factor. The authors used Collaborative filtration, Content-based as well as Hybrid methods to create a food and activity advice paradigm. A few issues with customer information, suggestion methodologies, altering food patterns, and set of data accessibility are mentioned as potential concerns. In 2019, Adrian B. R. Shatte et al. [25] concentrated their efforts on the usefulness of machine learning (ML) in the identification and treatment of behavioral problems like stress, Alzheimer's illness, and schizophrenic. There was also a rising desire while using the exploratory approval process to extend Machine Learning to other fields of research studies. More studies to find extra advantages of machine learning in these domains. Scientists and physicians will have easier access to machine learning tools. N. Yuvaraj et al. [26] propose a unique application of machine learning (ML) techniques for diabetic forecasting in Hadoop-based systems in 2019. Ultimately, when calculating the actual efficiency of the several machine learning (ML) methods, the Naive Bayes approach outperforms the DT (decision tree) technique by 3 percent. Research study can use Meta heuristic techniques as component of machine learning (ML) algorithms by adding additional units to Hadoop platform.

The conceptual framework for the phase forecasting of cervical cancer by Jaswinder Singh et al. [27] in 2019 allows wireless devices to capture health information, and the article's suggested framework clearly forecasts the proper phase for cervical cancer. In the future, one can enhance efficiency by taking into account additional factors such as a people's medical, ethnic, and dietary patterns. This work proposed the Naive Bayesian Classifier (i.e. RNBC) as a prognosis and therapeutic strategy for cardiovascular risk. The set of data is examined just once, offline, and the classification rule is designed without screening the set of statistics again. More classification strategies, including such decision

trees (DT), K nearest neighbor, and several other classification methods, might be performed and evaluated in further study to determine the much more accurate method. Abhaya Kumar Sahoo et al. [28] presented an HRS system that is based on analyzing huge data in 2019. It can anticipate ailments and gather information on unknown illnesses, allowing sufferers to be treated. The authors of these RSs used Hadoop and the Cassandra dataset. The CF method will also have substantial technical challenges as the percentage of subscribers and goods grows, and yet another key issue is that cold-start issues rise when the HRS would not have enough data on a specific physician or client to generate appropriate forecasts. According to Muhammad Waqar et al. [29], the article presents an adaptive method for successful development of physician suggestions. A study is used to identify certain physician characteristics. The program could be enhanced beyond by adding medications for patients and the indications of a certain condition. The suggested technology can be linked into any current hospital management system, enables individuals in urgent health - related issues to locate the proper clinician. In 2012, Asmaa S. Hussein et al. [30] used random forest (RF) algorithms in data mining to design an efficient and accurate CDD (Chronic Disease Diagnosis) recommendation systems solution. The approach utilized is likely to produce accurate illness risk predictions for chronically ill patients. Additional study into other prediction and recommendations approaches will be explored in terms of improving accurateness. As a result, the created model will be used to test more chronic illness research studies.

M. Janani et al. [31] provided suggestions for the aged depending on 3 primary terms in 2019. 1. Private details and interests of the elderly are collected and taught. 2. Medical problems including such heart rate and BP are assessed. 3 The mental state of the aged is determined by social contacts) the researcher used a Convolutional Neural Network based factor graph method which is an ontology system, a SVM with Artificial Neural Network, a RF (random forest), as well as a rule-based technique. The set of data of the worker strain advice method might be studied even more by researchers. Health Recommenders, developed by Abhaya Kumar Sahoo et al. [32] in 2019, are among the modern innovations used to retrieve relevant health - related information for a person. These algorithms determine favorite clinics by comparing the consistency of the decisions made by clients. As a result, they hold a crucial role in the healthcare system. RBM (Restricted Boltzmann Machine), Variable Weighted BSVD, DL, CNN methods were employed in the suggested RBM-Convolutional Neural Network. In the future, one could strive to improve the suggested method to improve greater performance while maintaining the highest level of anonymity.

There are typical applications of contagious illness prognosis focuses on ecological parameters including temperature (Hii Y.L et al., 2011) [34], (Lopman. B et al., 2009)[35]. Previous research has shown that meteorological data is a significant role in the spread of diseases (Huang X, Williams et al.) [36], (In 2015, Liu. T, Zhang et al.) [37], (In 2013, Blanford. J.I et al.) [38]. Precipitation and moisture are possible causes for viral disease with renal dysfunction, according to Liang et al. (Noden, B.H et al.)[39]. in addition, Huang et al. (In 2018, Liang. W et al.) [40] Observed that increases in dengue fever exhibit a high association with humidity levels. Earlier research has shown that using meteorological parameters, IOT, and Deep Learning (DL), infectious illness can be forecast more accurately.

3. PROPOSED SYSTEM DESIGN

A comprehensive explanation of the proposed system is shown in Figure 2, which may be seen below. There are a variety of methods that have been established in the past that may be used to detect the VBD that is mentioned in [15] and [6]. Yet, these kinds of systems continue to have problems, such as a high rate of false alarms and a low degree of accuracy in categorization. At the first step of the process, we gather information from a wide range of data sources. It is necessary to do this kind of preprocessing on the data using a certain sampling strategy and data filtering techniques. It was decided to use data filtering in order to get rid of cases that had been misclassified, and they went with a method of systematic sampling in order to separate the data. Through analysis, it should be able to give attribute identification and tokenization. Other natural language processing (NLP) procedures, such as stop word removal, lemmatization, and cleaning procedure are a part of natural language processing. Moreover, the system uses feature extraction methods that include co-occurrence correlation. During this election, a number of different quality criteria were used, and those qualities were carried over to the categorization system. Both the data training and the data testing procedures were carried out using a similar process. In order to categorise the complete network, we made use of a variety of different machine learning strategies. These strategies included SVM, RF, ANN, Naive Bayes, Adaboost, and others. With the suggested system, it is possible to recognise the VBD run time data analysis and eradicate it using a machine learning algorithm. Also, this system gives the maximum accuracy for both synthetic and real-time streaming data from a variety of sources. In addition to this, it has a low error rate and is able to operate on both homogenous and heterogenous datasets in a sequential manner.

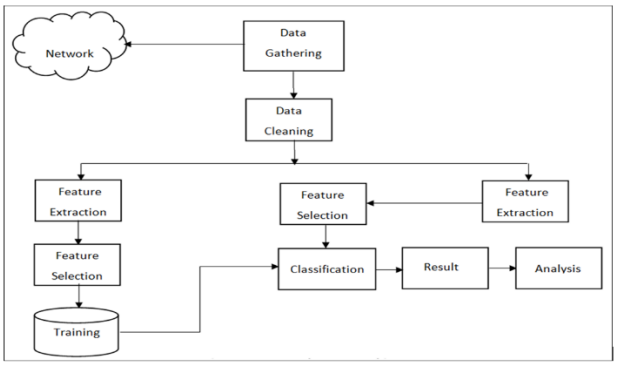

Figure 2: System architecture for VBD detection and classification using machine learning technique

Pre-processing:

Data has been validated according to the rules and norms defined during pre-processing. Each property has a lower and upper limit for specific values, and when one of these values exceeds or violates the bounds, the system instantly destroys the instance. Collection of data, gathering, cleaning, filtering and normalizing are all part of pre-processing.

Data Cleaning

Cleaning and repairing false or incorrect information from records, and datasets entails locating and changing as well as replacing, updating or sensitive information. They may use scripting software or transaction processing to sanitize data interactively. We used consistent sampling procedures to balancing data and filtered the standardized dataset to exclude the incorrectly classified occurrences.

Extraction of feature:

The normal and numerical values from the text data are extracted in the feature extraction step of the process. In order to remove characteristics such as TF-IDF, Co-relation co-occurrence, relational features, and dependency features from the complete dataset, a variety of approaches must be used to extract those features.

Feature Selection:

When feature extraction has been completed, the process of optimising the feature set using a small number of quality criteria is referred to as feature selection. A method called weighted term frequency has been used in order to optimise the features, and the results have been submitted to the training module.

Classification:

Finally, the system detects each record, either abnormal or normal using a supervised classification technique. Moreover, the system also demonstrates an neutral class classification of real time dataset. In this work, we carried out various supervised machine classification algorithm. Then supervised machine learning is applied in order to train the classifier. Here class labelled data is present at the beginning. The results have been evaluated according to confusion matrix and generated the precision, recall, accuracy, F1-Score etc in result section.

4. ALGORITHM DESIGN

Training Procedure

Input: Train dataset Training-Data [], several activation fun [], Threshold Th

Output: Extract Feat set[] for a trained unit that has been completed.

Step 1: Set the block input data dt[], the activation fun, and the epoch length.

Step 2: Feat-pckl ← Feat-Extract (dt [])

Step 3: Feat-set [] ← optimized (Feat-pckl)

Step 4: Return Feat-set []

System testing algorithm

Testing Process

Input: Normalized training dataset *Train_Data[]*, Normalized testing dataset *Test_Data[]*, defined threshold *qTh*

Output: Result set as output with *{Predicted_class, weight}*

Step 1: Read all test data from *Test_Data[]* using below function for validating to training rules, the data is normalized and transformed according to algorithms requirements

$$\text{test\_Feature(data)} = \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m] ... .A[n]} \leftarrow Test\_Data)$$

Step 2 : select the features from extracted attributes set of test_Feature(data) and generate feature map using below function.

$$\text{Test\_FeatureMap [t.....n]} = \sum_{x=1}^{n}(t) \leftarrow \text{test\_Feature(x)}$$

Test_FeatureMap [x] are the selected features in pooling layer. The convolutional layer extracts the features from input and passes to pooling layer and those selected features are stored in *Test_FeatureMap*

Step 3: Now read entire taring dataset to build the hidden layer for classification of entire test data in sense layer,

train_Feature(data)

$$= \sum_{m=1}^{n} (.\ \text{Attribute\_Set[A[m] ... . A[n]} \leftarrow Train\_Data)$$

Step 4 : Generate the training map using below function from input dataset

$$\text{Train\_FeatureMap [t.....n]} = \sum_{x=1}^{n}(t) \leftarrow \text{train\_Feature(x)}$$

Train_FeatureMap[t] is the hidden layer map that generates feature vector for build the hidden layer. That evaluate the entire test instances with train data.

Step 5 : After generating the feature map we calculate similarity weight for all instances in dense layer between selected features in pooling layer

Gen_weight

$$= \text{CalcWeight (Test\_FeatureMap} \,||\, \sum_{i=1}^{n} \text{Train\_FeatureMap[i])}$$

Step 6 : Evaluate the current weight with desired threshold

$$\text{if(Gen\_weight} >= \text{qTh)}$$

Step 7 : Out_List. add **(**trainF. class, weight**)**

Step 8 : Go to step 1 and continue when Test_Data == null

Step 9 : Return Out_List

5. RESULTS AND DISCUSSIONS

In a separate piece of research, using these assessment measures, it was shown that out of all the algorithms that were evaluated, our brand-new mSVM algorithm, which is also the name of the approach that we suggested, obtained the greatest predicted performance. In addition, due to the limitations imposed by the completeness and breadth of the data, many datasets may not accurately represent all of the real software problems and new defects. When implemented in real software applications, the solution that was provided may be subjected to further testing. It is possible to do 10, 15, and 20-fold cross-validation using the three data splitting technique.

Table 1: Dataset description of source code extracted from all data files

Total Size	10000
Training Samples	7000
Testing Samples	3000

The dataset is described in Table 1, which is located above, and a comparison study of the suggested model with the confusion matrix is included in Table 2, which is located below.

Table 1: Comparative analysis of proposed model with various machine learning classifiers

Methods	Precision	Recall	Accuracy	F-score
Naive Bayes	60.30	59.60	57.75	56.9
Adaboost	60.30	59.60	57.75	56.9
ANN	92.10	92.6	92.9	92.8
DT	55.10	56.3	55.90	56.00
SVM	98.50	97.3	99.08	98.50
mSVM	100	100	100	100

This system provides four comparisons between the findings of this study and the conclusions that some current systems have computed on different datasets that are comparable to those used in both sets. Figure 2 presents an analysis that contrasts newly discovered machine learning techniques with more traditional approaches.

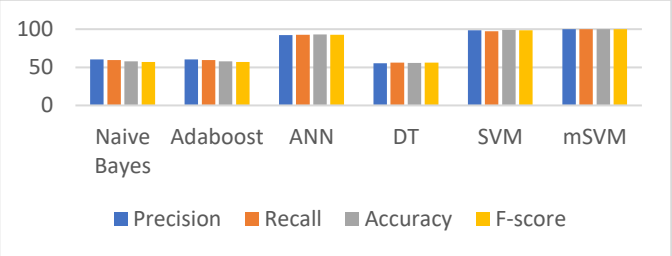

Figure 2: A comparison of the suggested classification with the current one for the purpose of VBD detection

Comparison of the classification accuracy of suggested machine learning methods for VBD detection with that of three current machine learning techniques is shown in figure 3 below. In terms of accuracy rate, this graphic demonstrates how the proposed mSVM performs better than other machine learning algorithms.

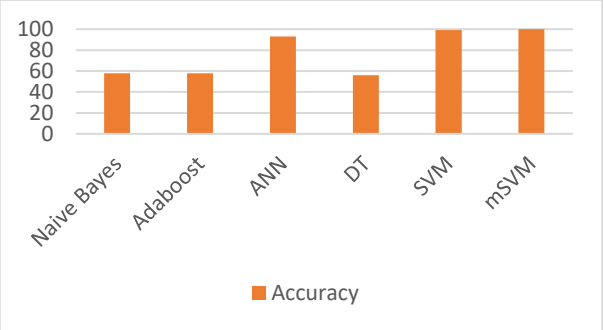

Figure 3: Analysis of the suggested classification in comparison to the current one for the identification of VBD

The classification accuracy of suggested methods for VBD detection is compared with that of four current machine learning techniques in figure 3, which can be found above. This chart illustrates how the proposed mSVM outperforms traditional machine learning techniques in terms of accuracy of detection.

6. CONCLUSION

This paper deals with the supervised machine learning approach to detection and classification Vector Borne Disease. The various supervised machine learning algorithms has used for achieving good accuracy. Comparing with existing machine learning algorithms our mSVM produces higher accuracy as 100% accuracy over the other machine learning classifiers. However, this study is limited patients' medical data but a better and more precise results could be obtained if consider the points such as large number of patients' data and more alternatives as well as criteria will be taken into consideration.

References

[1] Vijayakumar, V., Malathi, D., Subramaniyaswamy, V., Saravanan, P. and Logesh, R., 2019. Fog computing-based intelligent healthcare system for the detection and prevention of mosquito-borne diseases. Computers in Human Behavior, 100, pp.275-285.

[2] Inokuchi, M., Dumre, S. P., Mizukami, S., Tun, M. M. N., Kamel, M. G., Manh, D. H., ... & Hirayama, K. (2018). Association between dengue severity and plasma levels of denguespecific IgE and chymase. Archives of virology, 163(9), 2337-

[3] Iqbal N., & Islam, M. (2019). Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica, 43(3).

[4] Taneja, P., & Gautam, N. (2019). Hybrid Classification Method for Dengue Prediction. International Journal of Engineering and Advanced Technology (IJEAT).

[5] Guzman, M. G., & Kouri, G. (2003). Dengue and dengue hemorrhagic fever in the Americas: lessons and challenges. Journal of Clinical Virology, 27(1), 1-13.

[6] San Martín J. L., Brathwaite, O., Zambrano, B., Solórzano, J. O., Bouckenooghe, A., Dayan, G. H., & Guzmán, M. G. (2010). The epidemiology of dengue in the Americas over the last three decades: a worrisome reality. The American journal of tropical medicine and hygiene, 82(1), 128.

[7] Shepard, D. S., Undurraga, E. A., Betancourt-Cravioto, M., Guzman, M. G., Halstead, S. B., Harris, E., & Gubler, D. J. (2014). Approaches to refining estimates of global burden and economics of dengue. PLoS neglected tropical diseases, 8(11), e3306.

[8] Ibrahim F., Taib, M. N., Abas, W. A. B. W., Guan, C. C., & Sulaiman, S. (2005). A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN). Computer methods and programs in biomedicine, 79(3), 273-281.

[9] Gomes, A. L. V., Wee, L. J., Khan, A. M., Gil, L. H., Marques Jr, E. T., Calzavara-Silva, C. E., & Tan, T. W. (2010). Classification of dengue fever patients based on gene expression data using support vector machines. PloS one, 5(6), e11267.

[10] Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., ... & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. PLoS neglected tropical diseases, 11(10), e0005973.

[11] Carvajal, T. M., Viacrusis, K. M., Hernandez, L. F. T., Ho, H. T., Amalin, D. M., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC infectious diseases, 18(1), 1- 15.

[12] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, 16(3), 261-273.

[13] Indhumathi, K., & Kumar, K. S. (2021). A review on prediction of seasonal diseases based on climate change using big data. Materials Today: Proceedings, 37, 2648-2652.

[14] Alfred, R., & Obit, J. H. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. Heliyon, 7(6), e07371.

[15] Reddy, D. N. (2021). Machine Learning Algorithms for Detection: A Survey and Classification. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 3468-3475.

[16] alias Balamurugan, S. A., Mallick, M. M., & Chinthana, G. (2020). Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking. Informatics in Medicine Unlocked, 20, 100400.

[17] Ye, J., & Moreno-Madriñán, M. J. (2020). Comparing different spatio-temporal modeling methods in dengue fever data analysis in Colombia during 2012–2015. Spatial and Spatiotemporal Epidemiology, 34, 100360.

[18] Mussumeci, E., & Coelho, F. C. (2020). Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. Spatial and Spatio-temporal Epidemiology, 35, 100372.

[19] Chakraborty, T., Chattopadhyay, S., & Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. Physica A: Statistical Mechanics and its Applications, 527, 121266.

[20] Appice, A., Gel, Y. R., Iliev, I., Lyubchich, V., & Malerba, D. (2020). A multi-stage machine learning approach to predict dengue incidence: a case study in Mexico. Ieee Access, 8, 52713- 52725.

[21] Gambhir, S., Malik, S. K., & Kumar, Y. (2017). PSO-ANN based diagnostic model for the early detection of dengue disease. New Horizons in Translational Medicine, 4(1-4), 1-8.

[22] Mello-Román, J. D., Mello-Román, J. C., Gomez-Guerrero, S., & García-Torres, M. (2019). Predictive models for the medical diagnosis of dengue: a case study in Paraguay. Computational and mathematical methods in medicine, 2019.

[23] Portugal, I., Alencar, P. and Cowan, D., 2018. The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, pp.205-227.

[24] Vairale, V.S. and Shukla, S., 2019. Recommendation Framework for Diet and Exercise Based on Clinical Data: A Systematic Review. In Data Science and Big Data Analytics (pp. 333-346). Springer, Singapore.

[25] Shatte, A.B., Hutchinson, D.M. and Teague, S.J., 2019. Machine learning in mental health: a scoping review of methods and applications. Psychological medicine, 49(9), pp.1426-1448.

[26] Yuvaraj, N. and SriPreethaa, K.R., 2019. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing, 22(1), pp.1-9.

[27] Jaswinder Singh, Sandeep Sharma. 2019 Prediction of Cervical Cancer Using Machine Learning Techniques. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11 (2019) pp. 2570-2577.

[28] Sahoo, A.K., Pradhan, C., Barik, R.K. and Dubey, H., 2019. DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering. Computation, 7(2), p.25.

[29] Waqar, M., Majeed, N., Dawood, H., Daud, A. and Aljohani, N.R., 2019. An adaptive doctorrecommender system. Behaviour & Information Technology, pp.1

[30] Hussein, A.S., Omer, W., Li, X. and Ati, M., 2012. Accurate and reliable recommender system for chronic disease diagnosis. In GLOBAL HEALTH, The First International Conference on Global Health Challenges Venice, Italy.

[31] Janani, M. and Yuvaraj, N., 2019. Social Interaction and Stress-Based Recommendations for Elderly Healthcare Support System—A Survey. In Advances in Big Data and Cloud Computing (pp. 291-303). Springer, Singapore.

[32] Sahoo, A.K., Mallik, S., Pradhan, C., Mishra, B.S.P., Barik, R.K. and Das, H., 2019. Intelligence-Based Health Recommendation System Using Big Data Analytics. In Big Data Analytics for Intelligent Healthcare Management (pp. 227-246). Academic Press.

[33] Martínez-Pérez, B., De La Torre-Díez, I., & López-Coronado, M. (2015). Privacy and security in mobile health apps: a review and recommendations. Journal of medical systems, 39(1), 1-8.

[34] Hii Y.L, Rocklöv. J and Ng, N. Short Term Effects of Weather on Hand, Foot and Mouth Disease, PLoS ONE 2011, 6, e16796

[35] Lopman. B, Armstrong. B, Atchison. C and Gray, J.J. Host, Weather and Virological Factors Drive Norovirus Epidemiology: Time-Series Analysis of Laboratory Surveillance Data in England and Wales. PLoS ONE 2009, 4, e6671

[36] Huang X, Williams. G, Clements, A.C.A and Hu, W. Imported Dengue Cases, Weather Variation and Autochthonous Dengue Incidence in Cairns, Australia. PLoS ONE 2013, 8, e81887.

[37] Liu. T, Zhang. Y, Lin. H, Lv. X, Xiao. J, Zeng. W, Gu. Y, Rutherford. S, Tong S, Ma. W. A large temperature fluctuation may trigger an epidemic erythromelalgia outbreak in China. Sci. Rep. 2015, 5, 9525.

[38] Blanford. J.I, Blanford. S, Crane. R.G, Mann. M.E, Paaijmans. K.P, Schreiber. K.V, Thomas. M.B, Implications of temperature variation for malaria parasite development across Africa. Sci. Rep. 2013, 3, 1300.

[39] Noden, B.H.; Kent, M.D.; Beier, J.C. The impact of variations in temperature on early Plasmodium falciparum development in Anopheles stephensi. Parasitology 1995, 111, 539–545.

[40] Liang. W, Gu. X, Li. X, Zhang. K, Wu. K, Pang. M, Dong. J, Merrill. H.R,Hu. T, Liu, K; et al. Mapping the epidemic changes and risks of hemorrhagic fever with renal syndrome in Shaanxi Province, China, 2005–2016. Sci. Rep. 2018, 8, 749.