Online at: https://researchtrendsjournal.com **ISSN No:** 2584-282X
Indexed Journal **Peer Reviewed Journal**

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 3; Issue 6; 2025; Page No. 106-111

Received: 11-08-2025
Accepted: 23-09-2025
Published: 09-11-2025

Healthcare Data Analysis Using Competitive Ensemble Machine Learning Techniques

[1]Animesh Jain and [2]Dr. Bimal Kumar Rai

[1]Research Scholar, Department of Computer Science & Application, Mahakaushal University, Jabalpur, Madhya Pradesh, India
[2]Associate Professor, Department of Computer Science & Application, Mahakaushal University, Jabalpur, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.18265843

Corresponding Author: Animesh Jain

Abstract

The healthcare industry is rapidly adapting to the modern technological landscape, with many new advances appearing every year. This study aims to shed light on the Healthcare Data Analysis Using Competitive Ensemble Machine Learning Techniques Our goal is to present a concise outline of the many machine learning methodologies and to emphasize the domains where these methods are most commonly used, all while the healthcare business is seeing a wave of new machine learning applications. We discuss their prevalence and the potential for their further advancement in the medical field.

Keywords: Healthcare, Operations, Businesses, Medical, Illnesses

Introduction

Healthcare service Both quality and the ability to treat complex illnesses are constantly changing. Nevertheless, there are several challenges to overcome., especially when it comes to tailoring treatment regimens to specific patients or populations with limited access to clinical trials, like children. Consequently, ML has been effectively utilized in pediatric care for the purpose of predicting the optimal and tailored therapies for children in the past several years. The rapid spread of the COVID-19 epidemic has put ML in the limelight. In an often unpredictable and unpredictable work climate, organizations have embraced ML as a means to gain an edge and remain competitive. ML helps streamline operations and drives research and development. Hospitals and health systems have been able to tackle unique difficulties with the aid of ML.

Many businesses are trying to figure out how to use ML technology as it is among artificial intelligence's most intriguing subfields. The popularity of ML is skyrocketing. It has applications in both healthcare and industry and employs algorithms to make data-driven learning easier. As new ideas and technologies emerge, healthcare is also evolving at a rapid pace. In certain of these novel contexts, ML may be useful to healthcare providers. Though it was difficult to produce and use on a big scale in the past, modern technologies is now able to draw conclusions from unstructured text. With the abundance of new information made possible by machine learning, doctors and administrators may now make well-informed decisions about patient care and operational projects that impact the lives of millions of people in a timely manner.

The majority of ML issues are often well-constrained, meaning that there are usually no major obstacles to transforming the inputs into useful variables for model construction. Data pertaining to natural signals, noises, language, visual sceneries, or natural imagery is notoriously difficult to describe, yet it often appears in complicated real-world issues. This problem prompted the development of A new method for machine learning algorithms is called deep learning. Deep learning, in contrast to more traditional machine learning methods, works with a broad range of learning methods. Unlike the latter, which necessitates feature creation by domain experts, deep learning does feature engineering automatically, meaning it identifies relevant and informative features in the data before carrying out the specified job (e.g. classification, regression).

Literature and Review

Cruz *et al*. (2017) [1], have highlighted that so that computers may discover patterns from big, noisy, or complicated datasets, machine learning makes use of a number of optimization, statistical, and probabilistic methods. Their study touched on a few topics: cancer recurrence prediction, cancer survival prediction using a mixed-machine-learning technique, and cancer risk or susceptibility prediction utilizing several ML methods. They used Machine Learning techniques including ANN, SVM, Clustering, Genetic Algorithm, Decision tree, Naïve Bayes, and Fuzzy logic in their study on cancer prediction and the many kinds of cancers, including breast, skin, brain, cervical, colorectal, liver, lung, and throat cancers.

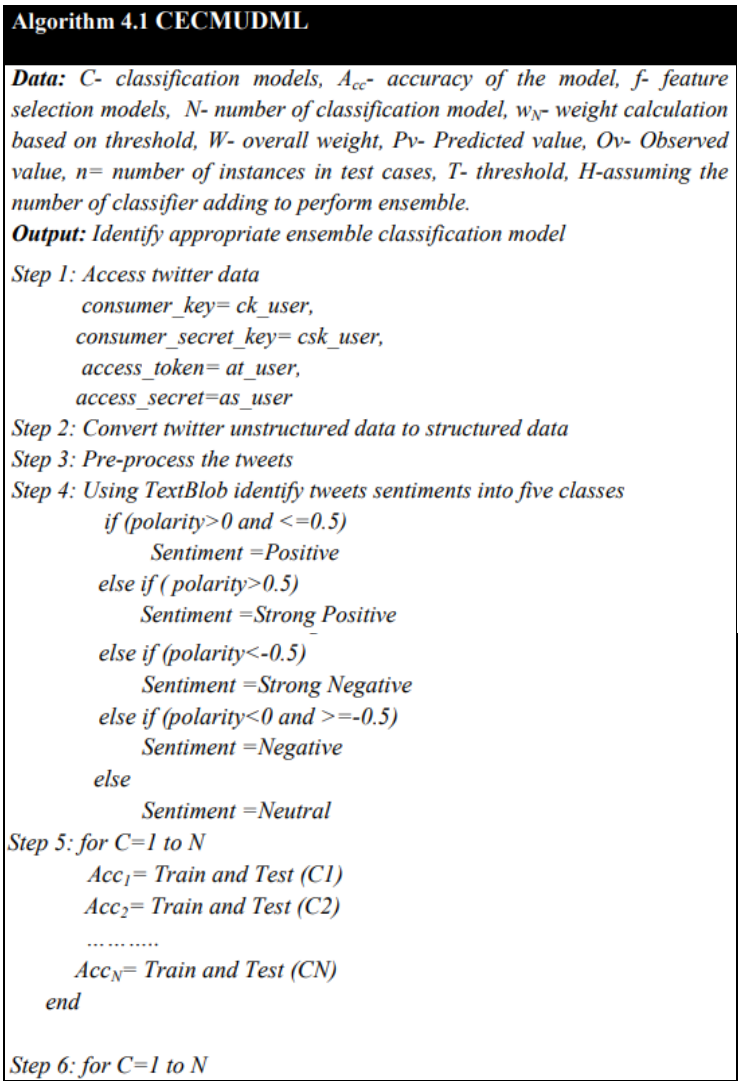
Ashfaq Ahmed *et al*. (2022) [2], Developed a Cancer Disease Prediction using Machine Learning classification methods such as Support Vector Machine and Random Forest to learn, classify, and compare data on cancer diseases utilizing various kernels and kernel parameters, including linear, polynomial, radial basis, and sigmoid. You can see that the classification accuracy may be changed by using a different kernel function and a different probabilistic estimate.

Nasser H. Sweilam *et al*. (2020) [3], A comparison research on support vector machines for cancer detection has been investigated by. Optimization via particle swarms, Q-behave particle swarms for SVM training, and Least squares SVM are the learning techniques that were presented in this study. We compare the outcomes of various approaches to the correct solution model issue and test them on a dataset pertaining to breast cancer.

Gayathri *et al*. (2023) [4], Findings from a survey on the use of ML algorithms such as SVM and RVM for breast cancer detection has been proposed by.

Parul Sinha *et al*. (2015) [5], When compared with Support Vector Machine (SVM), K-Nearest Neighbour (KNN) achieved better results in terms of accuracy, precision, and runtime comparative research on Chronic Kidney Disease Prediction by.

Proposed Using Machine Learning Techniques for Competitive Ensemble Classification of Unstructured Data (CECMUDML)



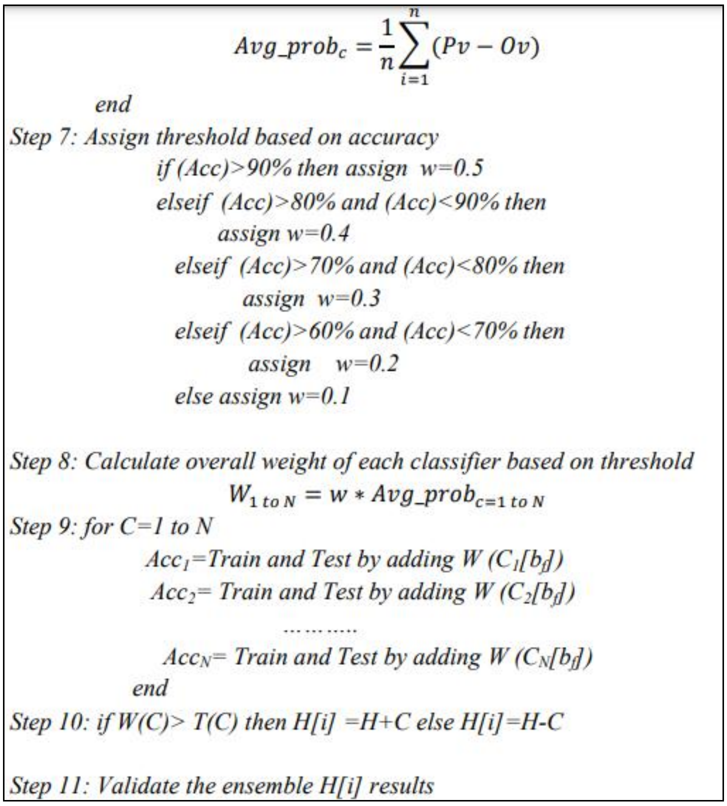
Algorithm 4.1 CECMUDML

Data: *C- classification models, A_{cc}- accuracy of the model, f- feature selection models, N- number of classification model, w_N- weight calculation based on threshold, W- overall weight, Pv- Predicted value, Ov- Observed value, n= number of instances in test cases, T- threshold, H-assuming the number of classifier adding to perform ensemble.*

Output: *Identify appropriate ensemble classification model*

```
Step 1: Access twitter data
        consumer_key= ck_user,
        consumer_secret_key= csk_user,
        access_token= at_user,
        access_secret=as_user
Step 2: Convert twitter unstructured data to structured data
Step 3: Pre-process the tweets
Step 4: Using TextBlob identify tweets sentiments into five classes
            if (polarity>0 and <=0.5)
                  Sentiment =Positive
            else if ( polarity>0.5)
                  Sentiment =Strong Positive
            else if (polarity<-0.5)
                  Sentiment =Strong Negative
            else if (polarity<0 and >=-0.5)
                  Sentiment =Negative
            else
                  Sentiment =Neutral
Step 5: for C=1 to N
        $Acc_1$= Train and Test (C1)
        $Acc_2$= Train and Test (C2)
        ………..
        $Acc_N$= Train and Test (CN)
    end

Step 6: for C=1 to N
```



```
                $Avg\_prob_c = \frac{1}{n}\sum_{i=1}^{n}(Pv - Ov)$
    end
Step 7: Assign threshold based on accuracy
            if (Acc)>90% then assign  w=0.5
            elseif  (Acc)>80% and (Acc)<90% then
                assign w=0.4
             elseif  (Acc)>70% and (Acc)<80% then
                  assign  w=0.3
             elseif  (Acc)>60% and (Acc)<70% then
                  assign    w=0.2
             else assign w=0.1

Step 8: Calculate overall weight of each classifier based on threshold
                $W_{1\ to\ N} = w * Avg\_prob_{c=1\ to\ N}$
Step 9: for C=1 to N
            $Acc_1$=Train and Test by adding W ($C_1[b_{fl}]$)
            $Acc_2$= Train and Test by adding W ($C_2[b_{fl}]$)
                  ………....
             $Acc_N$= Train and Test by adding W ($C_N[b_{fl}]$)
       end
Step 10: if W(C)> T(C) then H[i] =H+C else H[i]=H-C

Step 11: Validate the ensemble H[i] results
```

The CECMUDML consists of 11 steps described in algorithm 1. Step 1 represents the extraction of Twitter data using customer ID, customer secret key, data token, and data key. The second step is to use the established procedure to transform the unstructured data into the structured format. The third step explains how the tweets are prepared for processing. Section 4 describes the sentiment assignment for each tweet. Using TextBlob, the polarity rate for tweets is fixed based on a threshold. If "Strong positive" is the attitude acquired when the value is more than 0 but less than or equal to 0.5. In cases when the cutoff value exceeds half, then it acquires the sentiment of “positive”. Likewise, when the polarity value is below the threshold value as well as greater than -0.5, it acquires the sentiment of “strong negative”. If the polarity value lies between 0 and -0.5, it acquires sentiment as “negative”. If the polarity rate is zero, then its tweet sentiment is “neutral”. Step 5 designates the classification accuracy of each model represented as a variable named Acc_N. Step 6 assigns assigning a threshold value to each classifier in order to choose the ensemble's optimal performer. The seventh step is to give each classifier a weight according to its accuracy. In Step 8, the overall weight formula is described, where n is the number of rows, P is the classifier's projected value for the test data, and O is the observed value. In order to train the classifier using the correct weight from step 8, this weight is computed. Step 9 involves using each to train classifiers and then calculating their accuracy. At the same time, the competitive classifier is described in step 10. There is a weighted competition between each classifier; each classifier is trained and tested based on weight. Step 11 is to evaluate the results.

Results and Discussions

Table 1 describes the classification values for 5 sentiment labels. The Macro_Average (MA) and Weighted_Average (WA) for each classifier are calculated with the respective accuracy metrics. For the Diabetes Twitter data (Precision) *‘Pre’* value for the LR classifier, is high compared to other classifiers. (Recall) *‘Rec’* value the RF classifier's MA value is rather high, at 83% and a WA value of 95%. Likewise, the F1 score as F1_S is high for the RF classifier. Table 2 describes the Type 1 Diabetes dataset. Among all metrics, the LR classifier performs well compared to other classifiers. In Table 3, the SVM classifier performs well for ‘*Pre’* and (F1_ Score) ‘*F1_S’*. For ‘*Rec’*, the RF classifier performs better for Type 2 diabetes classification. Furthermore, Table 4 and Table 5 represent the Gestational Diabetes and Young Diabetes datasets. Both tables show the SVC classifier performs well for all metrics. In Table 6, the RF classifier performs better for all metrics for Diabetes Food tweets. Table 7 shows that the LR classifier performs better than all other classifiers for Diabetes Drug classification tweets.

Table 1: Diabetes Dataset Sentiment Classification Result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.94	0.99	0.96
	1	0.97	0.91	0.94
RF	2	0.94	0.80	0.87
	3	0.92	0.61	0.73
	4	1.00	0.63	0.91
	MA	0.95	0.83	0.88
	WA	0.95	0.95	0.95
	0	0.96	0.99	0.98
	1	0.96	0.95	0.96
LR	2	0.96	0.76	0.85
	3	0.97	0.41	0.58
	4	0.99	0.59	0.74
	MA	0.97	0.74	0.82
	WA	0.96	0.96	0.96
	0	0.86	1.00	0.93
	1	0.99	0.73	0.84
KN	2	0.98	0.59	0.60
	3	0.94	0.44	0.80
	4	1.00	0.67	0.80
	MA	0.96	0.69	0.78
	WA	0.91	0.90	0.89
	0	0.92	0.91	0.92
	1	0.83	0.93	0.88
NB	2	1.00	0.12	0.22
	3	0	0	0
	4	1.00	0.77	0.87
	MA	0.75	0.55	0.58
	WA	0.89	0.88	0.87
	0	0.87	0.98	0.92
	1	0.95	0.85	0.90
SVM	2	1.00	0.46	0.63
	3	1.00	0.18	0.31
	4	1.00	0.45	0.67
	MA	0.96	0.59	0.68
	WA	0.91	0.90	0.89

Table 2: Type 1 Diabetes Dataset Sentiment Classification Result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.93	0.98	0.95
	1	0.97	0.91	0.94
	2	0.97	0.81	0.88
RF	3	1.00	0.85	0.92
	4	1.00	0.91	0.95
	MA	0.97	0.89	0.93
	WA	0.95	0.95	0.95
	0	0.91	0.97	0.94
	1	0.94	0.92	0.93
	2	1.00	0.60	0.75
LR	3	1.00	0.08	0.14
	4	1.00	0.84	0.91
	MA	0.97	0.68	0.74
	WA	0.93	0.93	0.86
	0	0.76	1.00	0.79
	1	1.00	0.65	0.67
	2	1.00	0.50	0.38
KN	3	1.00	0.23	0.93
	4	1.00	0.86	0.83
	MA	0.95	0.65	0.72
	WA	0.87	0.83	0.82
	0	0.94	0.92	0.93
	1	0.82	0.95	0.88
	2	0.89	0.30	0.45
NB	3	0.92	0.03	0.07
	4	1.00	0.42	0.59
	MA	0.91	0.52	0.58
	WA	0.90	0.90	0.89
	0	0.94	0.95	0.96
	1	0.85	0.92	0.89
	2	0.90	0.25	0.40
SVM	3	1.00	0.03	0.05
	4	1.00	0.40	0.57
	MA	0.94	0.51	0.57
	WA	0.91	0.91	0.90

Table 3: Type 2 Diabetes Dataset Sentiment Classification Result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.88	0.98	0.93
	1	0.97	0.86	0.91
	2	0.97	0.70	0.81
RF	3	1.00	0.50	0.67
	4	1.00	0.94	0.97
	MA	0.96	0.80	0.86
	WA	0.92	0.92	0.92
	0	0.87	0.98	0.92
	1	0.95	0.87	0.91
	2	0.98	0.54	0.70
LR	3	1.00	0.33	0.50
	4	1.00	0.94	0.97
	MA	0.96	0.73	0.80
	WA	0.91	0.91	0.91
	0	0.74	1.00	0.85
	1	1.00	0.65	0.79
	2	0.91	0.49	0.64
KN	3	1.00	0.33	0.50
	4	1.00	0.94	0.97
	MA	0.93	0.68	0.75
	WA	0.86	0.82	0.81
	0	0.84	0.92	0.88
	1	0.87	0.86	0.86
	2	1.00	0.10	0.18
NB	3	0.00	0	0.00
	4	1.00	0.94	0.97
	MA	0.74	0.56	0.58
	WA	0.86	0.86	0.84
	0	0.90	0.98	0.94
	1	0.97	0.90	0.93
	2	0.98	0.70	0.82
SVM	3	1.00	0.33	0.50
	4	1.00	0.94	0.97
	MA	0.97	0.77	0.83
	WA	0.94	0.93	0.93

Table 4: Young Diabetes Dataset Sentiment Classification Result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.95	0.98	0.95
	1	0.97	0.95	0.94
	2	0.97	0.81	0.88
RF	3	1.00	0.85	0.92
	4	1.00	0.91	0.95
	MA	0.97	0.89	0.93
	WA	0.96	0.96	0.95
	0	0.98	0.94	0.96
	1	0.99	1.00	0.99
	2	1.00	0.82	0.90
LR	3	0.0	0	0.00
	4	0.0	0	0.00
	MA	0.59	0.55	0.57
	WA	0.98	0.99	0.98
	0	0.65	0.98	0.78
	1	1.00	0.93	0.96
	2	0.95	0.86	0.90
KN	3	0	0	0
	4	0	0	0
	MA	0.52	0.55	0.53
	WA	0.95	0.93	0.93
	0	0.65	0.96	0.78
	1	0.99	0.93	0.96
	2	1.00	0.70	0.82
NB	3	0	0	0
	4	0	0	0
	MA	0.53	0.52	0.51
	WA	0.95	0.93	0.93
	0	0.98	0.97	0.97
	1	0.99	1.00	0.97
	2	0.95	0.87	0.93
SVM	3	0.92	0.97	0.97
	4	0.98	0.97	0.95
	MA	0.98	0.97	0.98
	WA	0.97	0.98	0.98

Table 5: Gestational diabetes dataset sentiment classification result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.88	0.98	0.92
	1	0.94	0.85	0.89
	2	1.00	0.58	0.73
RF	3	1.00	0.45	0.62
	4	1.00	0.45	0.62
	MA	0.96	0.66	0.75
	WA	0.91	0.90	0.90
	0	0.87	0.97	0.92
	1	0.92	0.87	0.90
	2	1.00	0.31	0.47
LR	3	1.00	0.18	0.31
	4	1.00	0.09	0.17
	MA	0.96	0.48	0.55
	WA	0.90	0.89	0.88
	0	0.74	1.00	0.85
	1	1.00	0.60	0.75
	2	1.00	0.37	0.54
KN	3	1.00	0.18	0.31
	4	1.00	0.09	0.17
	MA	0.95	0.45	0.52
	WA	0.86	0.81	0.79
	0	0.87	0.87	0.87
	1	0.79	0.89	0.84
	2	1.00	0.15	0.27
NB	3	0	0	0
	4	0	0	0
	MA	0.53	0.38	0.39
	WA	0.82	0.83	0.82
	0	0.87	0.99	0.93
	1	0.97	0.83	0.89
	2	1.00	0.46	0.63
SVM	3	1.00	0.42	0.59
	4	1.00	0.64	0.78
	MA	0.97	0.67	0.76
	WA	0.91	0.90	0.90

Table 6: Diabetes Food Dataset Sentiment Classification Result

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.96	0.90	0.93
	1	0.93	0.99	0.96
	2	0.95	0.78	0.86
RF	3	0.96	0.57	0.72
	4	1.00	0.70	0.83
	MA	0.96	0.79	0.86
	WA	0.94	0.94	0.94
	0	0.96	0.91	0.93
	1	0.94	0.99	0.96
	2	0.95	0.77	0.85
LR	3	0.90	0.43	0.58
	4	1.00	0.61	0.76
	MA	0.95	0.74	0.82
	WA	0.94	0.94	0.94
	0	0.98	0.64	0.78
	1	0.82	0.99	0.90
	2	0.96	0.55	0.70
KN	3	0.89	0.43	0.58
	4	0.99	0.64	0.78
	MA	0.93	0.65	0.75
	WA	0.87	0.85	0.84
	0	0.69	0.94	0.80
	1	0.80	0.92	0.95
	2	1.00	0.71	0.84
NB	3	0	0	0
	4	0	0	0.70
	MA	0.53	0.52	0.51
	WA	0.96	0.93	0.93
	0	0.86	0.70	0.74
	1	0.70	0.86	0.75
	2	0.76	0.76	0.75
SVM	3	0.77	0.75	0.75
	4	0.86	0.65	0.74
	MA	0.66	0.86	0.75
	WA	0.86	0.80	0.74

Table 7: Diabetes Drug Dataset Sentiment Classification

Methods for Multiclass	Class	Pre	Rec	F1_S
	0	0.98	0.90	0.93
	1	0.93	0.99	0.96
	2	0.94	0.78	0.86
RF	3	0.96	0.57	0.72
	4	1.00	0.75	0.83
	MA	0.96	0.79	0.82
	WA	0.94	0.93	0.93
	0	0.98	0.93	0.93
	1	0.95	0.99	0.98
	2	0.96	0.80	0.85
LR	3	0.93	0.55	0.58
	4	1.00	0.61	0.76
	MA	0.96	0.74	0.82
	WA	0.97	0.94	0.94
	0	0.98	0.70	0.78
	1	0.82	0.99	0.90
	2	0.94	0.70	0.71
KN	3	0.90	0.43	0.58
	4	0.99	0.64	0.80
	MA	0.93	0.65	0.75
	WA	0.91	0.64	0.80
	0	0.65	0.96	0.78
	1	0.99	0.93	0.96
	2	1.00	0.70	0.82
NB	3	0	0	0
	4	0	0	0
	MA	0.53	0.52	0.51
	WA	0.95	0.93	0.93
	0	0.86	0.65	0.74
	1	0.66	0.86	0.75
	2	0.76	0.76	0.75
SVM	3	0.77	0.75	0.75
	4	0.86	0.65	0.74
	MA	0.66	0.86	0.75
	WA	0.80	0.81	0.75

Table 8: Twitter Classification using CECMUDML

Datasets	Method	Accuracy	Probability weight	Weight	Overall Weight	Individual accuracy	Ensemble Accuracy
Diabetes	MRFC	0.94	0.5	0.5	0.25	0.98	0.99
	MLR	0.96	0.5	0.5	0.25	0.96	
	KN	0.89	0.5	0.4	0.25	0.92	
	SVM	0.92	0.5	0.5	0.25	0.95	
	MNB	0.90	0.4	0.5	0.20 (Not)	-	
Type 1 Diabetes	RFC	0.94	0.2	0.5	0.1	0.95	0.96
	LR	0.92	0.2	0.5	0.1	0.93	
	KN	0.83	0.1	0.5	0.05(Not)	-	
	SVM	0.95	0.2	0.5	0.1	0.96	
	MNB	0.88	0.2	0.4	0.08(Not)	-	
Type 2 Diabetes	RFC	0.92	0.2	0.5	0.1	0.93	0.94
	LR	0.90	0.2	0.5	0.1	0.91	
	KN	0.82	0.2	0.4	0.08(Not)	-	
	SVM	0.93	0.2	0.5	0.1	0.94	
	MNB	0.85	0.2	0.4	0.08(Not)	-	
Young Diabetes	RFC	0.92	0.2	0.5	0.1	0.99	0.99
	LR	0.97	0.2	0.5	0.1	0.98	
	KN	0.95	0.2	0.5	0.1	0.97	
	SVM	0.95	0.2	0.5	0.1	0.99	
	MNB	0.96	0.2	0.5	0.1	0.97	
Gestational Diabetes	RFC	0.90	0.2	0.5	0.1	0.93	0.93
	LR	0.89	0.2	0.4	0.08(Not)	-	
	KN	0.80	0.2	0.4	0.08(Not)	-	
	SVM	0.90	0.2	0.5	0.1	0.92	
	MNB	0.83	0.2	0.4	0.08(Not)	-	
Diabetes Food	RFC	0.94	0.2	0.5	0.1	0.96	0.96
	LR	0.94	0.2	0.5	0.1	0.96	
	KN	0.85	0.2	0.4	0.08(Not)	-	
	SVM	0.80	0.2	0.4	0.08(Not)	-	
	MNB	0.81	0.2	0.4	0.08(Not)	-	
Diabetes	RFC	0.96	0.2	0.5	0.1	0.98	
Drugs	LR	0.95	0.2	0.5	0.1	0.97	
	KN	0.80	0.2	0.4	0.08(Not)	-	0.98
	SVM	0.85	0.2	0.4	0.08(Not)	-	
	MNB	0.82	0.2	0.4	0.08(Not)	-	

Conclusion

Additionally, the outcomes achieved via the suggested effort are contrasted with many conventional methods. The suggested solution outperformed established methods for forecasting chronic disease, according to the outcomes of the experiments. The proposed hybridized machine learning architecture, when combined with domain-specific feature selection, may effectively be used for chronic illness forecasting.

References

1. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informatics. 2017;16:59–77.
2. Ahmed KA, Aljahdali S, Hundewale N, Ahmed KI. Cancer disease prediction with support vector machine and random forest classification techniques. In: Proceedings of the IEEE International Conference on Computational Intelligence and Cybernetics (Cybernetics.com) Bali, Indonesia; c2012. p. 16–19.
3. Sweilam NH, Tharwat AA, Abdel Monie NK. Support vector machine for diagnosis of cancer disease: a comparative study. Egyptian Informatics Journal. 2020;11:81–92.
4. Gayathri BM, Sumathi CP, Santhanam T. Breast cancer diagnosis using machine learning algorithms: a survey. International Journal of Distributed and Parallel Systems. 2023;4(3):1–10.
5. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology. 2015;4(12):1–6.