International Journal of Engineering &Technology, 13 (1) (2024) 182-188

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Study and analysis of feature selection problems and impact of bias in machine learning disease prediction models

Anil Kumar Prajapati [1] *, Umesh Kumar Singh [1], Rekha Singh [2], Arpita Shukla [3]

[1] Institute of Computer Science, Vikram University Ujjain (MP)
[2] School of Computer Science & Information Technology, DAVV Indore (MP)
[3] JNS Govt. PG College Shujalpur, Shajapur (MP)
**Corresponding author E-mail: anilprajapatiujjain@gmail.com*

Abstract

In the current scenario machine learning is the branch of artificial intelligence being used in every field and medicine is one of them. In medical science, the use of machine learning techniques aims to improve patient care by collecting, and analyzing patient data, and designing advanced and intelligent tools and/or devices for disease detection using collective experience. ML technology detects patterns associated with specific diseases by analyzing large datasets that include various patient records, such as diabetes, blood pressure, cholesterol, X-rays, MRIs, CT scans, imaging data, and genomic information. ML algorithms compute the primary symptoms of the disease. Based on these calculations the disease is identified. Here it is necessary to have sufficient dataset and/or features for computation. The understanding of the ML model depends on the underlying feature to be used to identify the related problem. The fairness of a machine learning algorithm depends on which symptoms are selected to determine any disease. The selection of features for ML models is an important task, more or less features can make the model underfit or overfit. Incorrect determination of selected features can introduce bias into the model which can greatly affect the accuracy of the model. If the bias in the machine learning model is not properly tuned or the bias is tuned too high or too low then the prediction does not cover the underlined pattern. Diseases arise in different circumstances; each disease has its special characteristics. To cover all the basic parameters of each disease is a very tough task. If a basic attribute is missed and/or an attribute that has no relation to the disease is captured then the desired result of the model may be affected. In the proposed research paper, the feature selection problem and bias effect have been analyzed through the Support Vector Machine (SVM) and Logistic Regression (LR) algorithm.

Keywords: *Bias; Classification; Health Care; Fuzzy Logic; Machine Learning.*

1. Introduction

In recent years machine learning technology has had high attention in every field such as Pattern Recognition, Image Recognition, and Natural Language Processing. A major challenge in medicine has been determining disease risk, making a diagnosis, predicting treatment outcomes, or establishing a prognosis. Nevertheless, models are being developed to identify the disease by understanding the data patterns through machine learning algorithms, but the hidden values in the data and the maximum distance of symptoms of diseases from diseases can prove to be a vital threat. Machine learning techniques provide high-quality results due to accurate observations and high computing power [1-2-3]. ML technique works on regression and classification problems, there is a wide range of regression and classification algorithms available in ML technique. Disease prediction and/or detection come under classification issues in machine learning techniques. The phenomenon of classification depends on yes or no, it means diseases are present and absent. The symptoms and/or features used in machine learning algorithms to detect the disease play a significant role in prognosis. The outcome of a machine learning model depends on the data and algorithm used in the algorithm. The growing popularity of these techniques raised questions about biases embedded in them and about how fair these models are when defining their performance concerning sensitive social issues such as genetics, healthcare, and disease detection. Therefore, it is necessary to use the most consistent features related to the disease and avoid noisy features. Fuzzy logic can be used to determine feature weights and/or values used for ML model training, so that bias can be reduced or minimized. In this paper, we will use the Kaggle dataset of the 13 most valuable and updated features (Table 1.0) of cardiovascular disease, which is freely available on the Kaggle data repository for practical and testing purposes [4]. Logistic Regression and Support Vector Machine algorithms will be used to test the accuracy scores of different combinations of features. The primary goal of this paper is to examine the extent to which bias can impact human understanding and functionality of machine learning models when the data used is highly consistent and correlated.

2. Literature review

The landscape of medical science has changed to a great extent after the COVID-19 pandemic. Researchers and doctors regularly analyze and discover new patterns of disease diagnosis and/or detection. Technology enhances the capacity of medical science, hence the burden and/or responsibility of effectively identifying and/or detecting the disease remains a major concern and responsibility for the upcoming algorithm or technology [5-6]. The generalization of the disease depends on how much and effectively the symptoms related to the disease have been studied in the past. For accurate identification of the disease, the type of disease is determined by the doctors and researchers based on the symptoms related to the disease and the serious consequences caused by the disease. For accurate identification of the disease, the type of disease is determined by doctors and researchers based on the symptoms related to the disease and the serious consequences caused by the disease. In this way, the disease is determined based on certain criteria. Machine learning algorithms for disease detection determine disease based on certain parameters (symptoms) similar to human understanding [7-8]. Here it becomes important how the data related to the disease has been determined. Here the role of bias in the data becomes important, the success of the learning algorithm will depend on the bias determining the data type. Bias in learning algorithms is determined in many ways [9], [10], [11]. The development of ML-based medical applications can be biased across multiple steps in the process, such as data collection, data preparation, modeling, evaluating, and deploying the system into clinical practice after approval. During the machine learning process, algorithmic bias can be classified based on when it occurs. An unbalanced model is usually the result of biased training samples. Depending on the application, training data bias can originate from a variety of sources, such as human labeling, sample selection, and others. When algorithm output is interpreted by users, a post-algorithmic bias arises [12-13]. The fairness of any kind of ML model can be depicted in the absence and /or triviality of bias in the use of the training dataset. Bias represents the difference between the predicted output of the model and the true or expected output. Bias can manifest in various ways, and it is often associated with the model's inability to accurately capture the underlying patterns in the data [14-15]. An ML model is trained on fewer features of related problems then the model provides the wrong prediction, commonly this is known as a low-bias problem. While an ML model is trained on extra features of related problems the model provides the wrong prediction is known as a high-bias problem [16-17]. The major reason for bias occurring in the ML model is the maximum distance between features and disease or in other words, the feature to be used does not cover the underlying pattern of disease and hidden values [18]. Bias is classified into three major categories covariate shift bias, sample selection bias, and imbalance bias.

Covariate shift: In the Machine learning Model when the distribution of input features in the training data differs from the distribution of the features in testing data then a Covariate shift bias occurs. In other words when one of the important features is not covered in the testing datasets then the Covariate shift bias arises. The major impact of this bias is the model may struggle to generalize well to hidden patterns.

Sample selection bias: Sample selection bias depends on how sophisticated the samples selected for model training are. If there is no correlation between an event that occurred in the past and the current sample, the result will not be effective. Previous events will have a huge impact on the outcome. Here it would be effective to see how much correlation there is between the current sample and the previous and/or past sample.

Imbalance bias: Imbalance bias occurs when the data used to train a machine learning model is not uniform. That means some classes have significantly fewer instances used than others. Machine learning models can suffer from this class imbalance, especially in binary classifications. These are the fundamental types of bias in ML and DL models [19], [20], [21].

3. Impact and problem of bias

The fairness of the ML model will depend on features that are related to the disease. In some specific diseases like cardiovascular, cancer, tumors, paralysis, and brain hemorrhage there is a wide range of disease symptoms, hence it becomes very difficult to determine which symptoms will be used to build an accurate ML model. Machine learning techniques generate a prediction of disease based on previous records and some specific symptoms of the disease. The machine learning algorithm captures the relation between the disease and symptoms of diseases and generates the result [22]. The fairness of algorithms depends on how much the disease symptoms are nearest to the disease and how many symptoms used in model training. This problem is commonly known as the bias problem in machine learning models. Failure of the model to understand the hidden values and underlying patterns of the data or not achieving the proposed results indicates the presence of bias. Here it will be important to understand how and to what extent the accuracy of the model will be affected due to bias. When a machine learning model has a high bias in its datasets, the model will suffer from the problem of underfitting. Conversely, if there is low bias, the model will suffer from the problem of overfitting [23 - 26]. In previously proposed models it has been observed that increasing or decreasing the features used in model training has a great impact on the accuracy of the model. Ghosh et al. Relief and Absolute Shrinkage and Selection Operator (LASSO) techniques were used for feature extraction in model training. They achieved 89% accuracy using the first 13 features and achieved 92% and 99.05% accuracy on reducing the features i.e. using 11 and 10 features respectively [27]. It is noteworthy that accuracy is higher when features are fewer. The boundaries of the ML model show the limitations of the predictions. Hence the issues of developing an effective ML prediction model are as follows:

- How to minimize the bias in model training for optimum machine learning model
- How many attributes are taken for an optimum ML disease prediction model?
- How to decide, which attributes are closest to related disease?
- How to identify which attributes do not have hidden values that will never affect the prediction model?

4. Methodology

The problem of feature selection and bias in machine learning classification models is not new. In the proposed study we have used 1025 patient records with 13 different and important features of cardiovascular disease [4] that are listed in Table 1.0. In this approach to identify bias and feature selection problems, we used minimal cardiovascular features for cardiovascular disease detection and/or prediction.

Table 1: Dataset Description and Used Parameter

Sl. No.	Attributes	Description	Measurement Unit
1	Age	Age of the patients	In year
2	Sex	Sex of the patients	0,1(0 = Female, 1 = Male)
3	CP	Chest pain	0,1 (0 = Typical angina, 1= Atypical angina)
4	Trestbps	Resting Blood Pressure	94-200 (In mmHg)

5	Chol	Cholesterol	126-564 (in mg/dl)
6	Fbs	Fasting Blood Sugar	0,1 > 120mg/dl (0 = false, 1 = true)
7	Restecg	Resting ECG	The numerical value (0,1,2,)
8	Thalach	Max Heart Rate	71-202
9	Exang	Exercise Enigma	0,1 (0 = No, 1 = Yes)
10	Old peak	Old peak	0-6.2
11	Slop	Slop	1,2,3 (1 = up sloping, 2 = flat, 3 = down sloping)
12	Ca	No. of Major Vessels	0,1,2,3,
13	Thal	Thalassemia Display	3 = Normal, 6 = Fixed, 7 = Reversible defect

To detect bias in ML disease detection models and understand the importance of features, we divided the entire dataset into four different smaller datasets containing 1025 examples and 13 features. Out of 13 features, we randomly reduced features 1, 2, 3, and 4 and divided them into 12, 11, 10, and 09 features respectively. The details of the dataset created for feature selection and analysis of the bias problem are as follows.

a) A total of 13 tables were created with 12 attributes by subtracting 1 attribute from the total 13 attributes.
b) A total of 12 tables were created with 11 attributes by subtracting 2 attributes from the total 13 attributes.
c) A total of 07 tables were created with 10 attributes by subtracting 3 attributes from the total 13 attributes.
d) A total of 06 tables were created with 09 attributes by subtracting 4 attributes from the total 13 attributes.

The created datasets are divided 8:2 and 7:3 ration into training and testing datasets. On this training and test data, the support vector machines and logistic regression algorithms were used to make predictions. The overall work process of the proposed research work is shown in below figure 1.0

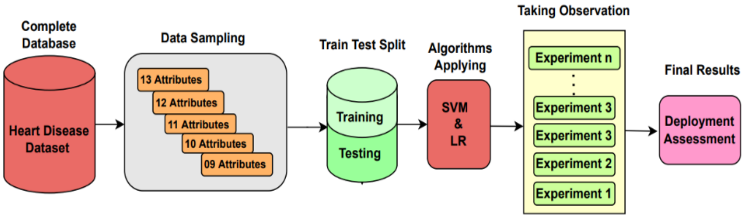

Fig. 1: ML Model for Disease Prediction.

SVM: Support vector machine is used in supervised learning for solving classification as well as regression problems. The algorithm works on classifying datasets into two categories that are positive and negative and/or present or absent. The algorithm develops a hyperplane and works to define boundaries between data points based on predefined classes, labels, or output [28-29].

LR: Logistic regression algorithms work on classification and/or binary classification. this is a statistical algorithm used to analyze the relationship between two factors in data. The algorithm uses the sigmoid function to linearly combine features and parameters and the sigmoid function maps any real-valued number to a value between 0 and 1[30-31]. To check the accuracy of the model we calculated the accuracy scores on different features using Python language on Jupyter Notebook of Anaconda Navigator. Support vector machines and logistic regression algorithms were used for overall calculations. The proposed research is based on a classification problem and both ML algorithms are used to solve the classification problem. The analysis of accuracy scores of the Support Vector Machine and Logistic Regression algorithm on tables created with different specifications is shown in Tables 2.0, 3.0, 4.0, and 5.0 below.

Table 2: Accuracy Scores on SVM and Logistic Regression Algorithm with 13 Features of Heart Disease

Sl. No.	13 Attributes of Heart Disease	SVM Accuracy Score		Logistic Regression Accuracy Score	
		Ratio 8:2	Ratio 7:3	Ratio 8:2	Ratio 7:3
1	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.95	87.02	85.24	86.47

Table 3: Accuracy Scores on SVM And Logistic Regression Algorithm with 12 Features of Heart Disease

Sl. No.	Combination of Attributes	Combination of 12 Attributes of Heart Disease	SVM Accuracy Score		Logistic Regression Accuracy Score	
			8:2	7:3	8:2	7:3
1	A	Sex, cp, Trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.34	87.02	84.87	86.47
2	B	Age, cp, Trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	85.24	85.63	85.12	85.21
3	C	Age, sex, Trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	82.02	82.98	82.92	84.51
4	D	Age, sex, cp, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.21	87.16	85.24	87.44
5	E	Age, sex, cp, Trestbps, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.34	87.72	85.85	86.75
6	F	Age, sex, cp, Trestbps, chol, restecg, thalach, exang, oldpeak, slope, ca, thal	87.19	86.75	87.24	87.88
7	G	Age, sex, cp, Trestbps, chol, fbs, thalach, exang, oldpeak, slope, ca, thal	86.82	87.72	86.21	87.16
8	H	Age, sex, cp, Trestbps, chol, fbs, restecg, exang, oldpeak, slope, ca, thal	85.48	86.33	85.60	86.75
9	I	Age, sex, cp, Trestbps, chol, fbs, restecg, thalach, oldpeak, slope, ca, thal	86.95	87.72	86.95	88.84
10	J	Age, sex, cp, Trestbps, chol, fbs, restecg, thalach, exang, slope, ca, thal	84.82	85.35	85.36	85.91

11	K	Age, sex, cp, Trestbps, chol, fbs, restecg, thalach, exang oldpeak ca thal	85.48	87.16	85.73	86.91
12	L	Age, sex, cp, Trestbps, chol, fbs, restecg, thalach, exang oldpeak slope thal	84.26	85.63	85.60	86.33
13	M	Age, sex, cp, Trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca,	84.14	85.77	85.48	86.47

Table 4: Accuracy Scores on SVM And Logistic Regression Algorithm with 11 Features of Heart Disease

Sl. No.	Combination of Attributes	Combination of 11 Attributes of Heart Disease	SVM Accuracy Score		Logistic Regression Accuracy Score	
			8:2	7:3	8:2	7:3
1	A	cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	85.00	85.63	84.26	84.93
2	B	Age, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	81.82	83.13	81.46	83.54
3	C	Age, sex, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	82.19	83.96	83.53	84.51
4	D	Age, sex, cp, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.36	86.50	85.99	87.16
5	E	Age, sex, cp, trestbps, restecg, thalach, exang, oldpeak, slope, ca, thal	85.97	86.47	85.00	87.17
6	F	Age, sex, cp, trestbps, chol, thalach, exang, oldpeak, slope, ca, thal	86.58	87.44	86.34	87.02
7	G	Age, sex, cp, trestbps, chol, fbs, exang, oldpeak, slope, ca, thal	84.39	85.77	81.46	84.65
8	H	Age, sex, cp, trestbps, chol, fbs, restecg, oldpeak, slope, ca, thal	84.75	85.91	82.19	85.49
9	I	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, slope, ca, thal	85.97	85.91	85.48	85.91
10	J	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, ca, thal	84.14	85.21	82.80	84.23
11	K	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, thal	84.87	85.21	84.46	85.49
12	L	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope	82.43	83.96	80.73	81.02

Table 5: Accuracy Scores on SVM and Logistic Regression Algorithm with 10 Features of Heart Disease

Sl. No.	Combination of Attributes	Description of 10 selected Attributes	SVM Accuracy Score		Logistic Regression Accuracy Score	
			8:2	7:3	8:2	7:3
1	A	Age, CP, Trestbps, Chol, Fbs, Thalach, Old peak, Slop, Ca, Thal	83.65	85.49	84.63	84.79
2	B	Age, CP, Trestbps, Chol, Fbs, Thalach, Exang, Old peak, Slop, Thal	82.80	84.51	81.70	82.56
3	C	Age, CP, Trestbps, Chol, Thalach, Exang, Old peak, Slope, Ca, Thal	87.90	88.77	88.20	88.99
4	D	Age, Sex, Trestbps, Chol, Fbs, Thalach, Exang, Old peak, Slop, Thal	82.07	84.00	81.75	82.42
5	E	Age, Sex, CP, Chol, Fbs, Restecg, Thalach, Exang, Ca, Thal	84.14	84.10	83.78	84.79
6	F	Age, Sex, Trestbps, Chol, Fbs, Thalach, Exang, Old peak, Slop, Ca	81.46	81.86	81.58	81.72
7	G	Age, Sex, CP, Trestbps, Fbs, Restecg, Thalach, Exang, Ca, Thal	83.51	85.07	83.78	85.49

Table 6: Accuracy Scores on SVM and Logistic Regression Algorithm with 09 Features of Heart Disease

Sl. No.	Combination of Attributes	Combination of 09 Attributes of Heart Disease	SVM Accuracy Score		Logistic Regression Accuracy Score	
			8:2	7:3	8:2	7:3
1	A	Age, chol, fbs, restec, thalach, exang, oldpeak, slope, ca,	81.16	81.72	81.21	82.84
2	B	Age, cp, trestbps, thalach, exang, oldpeak, slope, ca, thal	86.58	86.72	86.46	86.86
3	C	Age, sex, cp, trestbps, chol, fbs, restecg, slope, thal	82.80	83.96	82.19	82.70
4	D	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, oldpeak	82.43	83.54	81.46	81.86
5	E	Sex, cp, trestbps, restecg, thalach, oldpeak, slope, ca, thal	86.35	86.70	86.26	86.00
6	F	Age, cp trestbps, chol, restecg, thalach, exang, ca, thal	83.39	85.49	83.65	84.65

5. Result and discussion

Symptoms play an effective role in detecting the disease. There can be many symptoms related to the disease, and it is very important to recognize the effective symptoms, only then the disease can be accurately identified. In the proposed study, we took the 13 most relevant features of cardiovascular disease and identified CVD using a support vector machine and logistic regression algorithm, which achieved an accuracy of 87% and 86%, respectively. In the study, we observed that the maximum accuracy of the algorithm on 12 features after reducing one (fasting blood sugar) was 87.75 and 87.88. While reducing two features (fasting blood sugar and resting ECG) made no difference in the maximum accuracy of the algorithm (i.e. 87.44 and 87.02). The accuracy was also checked on 10 and 9 features respectively to understand the impact of bias and features. We observed that by reducing 3 features (i.e. sex, fasting blood sugar, and resting ECG), the accuracy of the model increases to 88.77 and 88.99 respectively, while by reducing 4 features (i.e. sex, cholesterol, fasting blood sugar, and resting ECG) The maximum accuracy of the model was observed to be 86.72 and 86.86. Table 6.0 and Graph 1.0 show the observations taken.

Table 7: Comparison of Different Combinations of Features and Highest Accuracy Score

Sl. No.	Combination of Attributes	Comparison of Selected Attributes	SVM Accuracy Score		Logistic Regression Accuracy Score	
			8:2	7:3	8:2	7:3
1	13	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal	86.95	87.02	85.24	86.47
2	12	Age, sex, cp, Trestbps, chol, restecg, thalach, exang, oldpeak, slope, ca, thal	87.19	86.75	87.24	87.88
3	11	Age, sex, cp, trestbps, chol, thalach, exang, oldpeak, slope, ca, thal	86.58	87.44	86.34	87.02
4	10	Age, CP, Trestbps, Chol, Thalach, Exang, Old peak, Slop, Ca, Thal	87.90	88.47	88.20	88.99
5	09	Age, cp, trestbps, thalach, exang, oldpeak, slope, ca, thal	86.58	86.72	86.46	86.86

The ML model achieved maximum accuracy on 10 different cardiovascular features that are Age, Chest pain, Resting Blood Pressure, Cholesterol, Max Heart Rate, Exercise Enigma, Old peak, Slop, No. of Major Vessels, and Thalassemia Display while sex, fasting blood sugar, and resting ECG are also played an important role in determining cardiovascular disease. Now here it is very important to decide which features should be removed in model training and which should not. The use of different features affected the accuracy of the model, now it is difficult to determine how many features should be used. Also, it becomes difficult to determine which features have a bias or not. Table 6.0 and Graph 1.0 show the observations taken. It is clear from the graph that the model achieved maximum accuracy on 10 features for heart disease. Minimizing bias and determining the important features of the disease is essential for an accurate and effective model design. Now the main problem that arises here is that if the model is trained on limited features, then the bias in the model will increase and the model will become underfit. Therefore, to obtain maximum accuracy from the model, we cannot reduce disease-related features in model training. So, the question is how many optimal features should be taken so that the effect of bias is minimal and the model is not underfitting. On the contrary, to reduce the bias, more features should be used but, in this situation, the model will be overfit and will provide wrong results. Hence the optimal features should be used necessary in model training so that the model provides effective prediction.

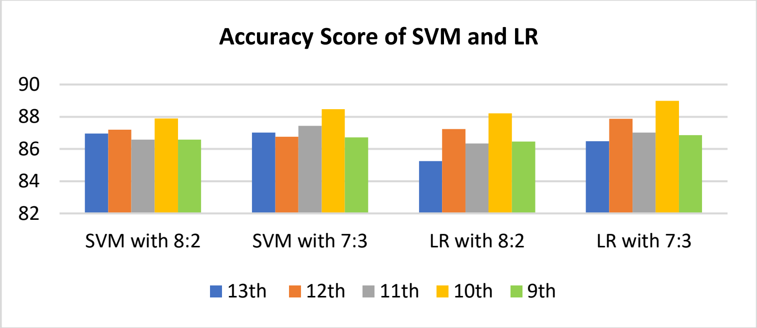

Fig. 2: Accuracy Score of SVM and LR.

6. Remedies

The results obtained from the Support Vector Machine and Logistic Regression algorithm show that bias and feature selection are major issues in machine learning models. If the features selected for model training have a maximum distance from the disease, the model will suffer from the problem of bias. Selecting optimal features in model training can prevent bias in the model so that the accuracy of the model will be favorable to the desired outcome. We would recommend using fuzzy logic for feature selection in machine learning model training. Fuzzy logic is capable of determining the values (hidden values) and boundaries of any features used in a learning model [32-33]. We can understand this as an example, if someone is having chest pain then three conditions will arise - very high pain, very low pain, low pain and/or very high blood pressure, high blood pressure, and only blood pressure. Here it is important to decide which data should be taken for model training. Determining this member function will decide which feature has maximum utility for model training. Member functions operate on mathematical notation which is as follows

Core: Region of the universe that is characterized by full membership in the set

$$\boldsymbol{\mu\ A\tilde{}\ (y) = 1.}$$

Support: Region of the universe that is characterized by a non-zero membership in the set

$$\boldsymbol{\mu A\tilde{}\ (y) > 0.}$$

Boundary: Region of the universe that is characterized by a non-zero but incomplete membership in the set

$$\boldsymbol{1 > \mu A\tilde{}\ (y) >.}$$

Diseases can have many symptoms and features; it is important to determine among the symptoms those that uniquely identify the disease. The values of features can be determined using the core, support, and boundary functions of fuzzy logic. We can decide the relevance of the feature to the respective disease on the behalf of member function which is to provide optimized solutions for feature selection and bias problems. This is a hypothetical approach to the related issues and model design. Only the experimental observations made will accurately verify the problem determination.

7. Conclusion

In the present context, increased mortality due to diseases is a matter of concern. Machines used in medical science have made disease diagnosis accessible to a great extent. In medical science, techniques like machine learning/deep learning can prove to be a better option for disease diagnosis and/or identification. In such a situation, making an effective machine that can accurately identify diseases is a very responsible task that is directly related to human life. The features used to identify the disease play a major role in the ML model; the ML model determines the disease based on features. In the proposed research, we tested the accuracy of the cardiovascular disease model on various features and found that the model gives high accuracy when fewer features are used, while the accuracy is lower when more features are used. This problem has been shown in the research that on what basis should the features used for model training be determined so that the problem of bias does not arise in the model. In the research, maximum accuracy was achieved by using 10 features of cardiovascular disease, but features like sex, fasting blood sugar, and resting ECG also cannot be ignored. If these are ignored and the focus is placed on maximum accuracy then the model may be affected by bias and if they are used then the accuracy is affected. In such a situation, it is important to solve the question of how many features should be selected and which feature should be given priority in selection. In the proposed research, we have presented the hypothetical idea of using fuzzy logic for feature determination, which can be worked on in the future. However, machine learning and/or deep learning are effective tools that are being used in medical science and this technology will yield more effective and sophisticated results in the future.

Acknowledgment

I express my gratitude to Dr. Umesh Kumar Singh, Director and Supervisor of the Institute of Computer Science, for his direct and indirect contribution to completing the article. Under whose guidance I did my research work.

References

[1] Li H, Deep learning for natural language processing: advantages and challenges, National Science Review, Volume 5, Issue 1, January 2018, Pages 24–26, https://doi.org/10.1093/nsr/nwx110.

[2] Bishop CM, “Neural Networks for Pattern Recognition”, Oxford University Press, Inc., USA, 1995. https://books.google.co.in/books?. https://doi.org/10.1093/oso/9780198538493.001.0001.

[3] Cios KJ, and Shin I, “Image recognition neural network: IRNN”, Neurocomputing 7 (1995) 159–185, https://doi.org/10.1016/0925-2312(93)E0062-I.

[4] Kaggle Heart Disease Dataset, Access date 15/10/2023 https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[5] Prajapati AK, and Singh UK, “Cardiovascular disease (CVD) prediction through Artificial Neural network in the perspective of Deep Learning”, International Journal of Computing Algorithm, Volume 11, Issue 2, 2022, pp. 1-7, https://doi.org/10.20894/IJCOA.101.011.002.002.

[6] Alanazi R, “Identification and Prediction of Chronic Diseases Using Machine Learning Approach”, Journal of Healthcare and Engineering, Volume 2022, PP. 2022:2826127. Feb - 2022, https://doi.org/10.1155/2022/2826127.

[7] Kliegr T, Bahnik S, Furnkranz J, “A review of possible effects of cognitive biases on interpretation of rule-based machine learning models”, Artificial Intelligence, Volume 295,2021, PP.103458, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2021.103458.

[8] Prajapati AK and Singh UK, “An empirical analysis of ML techniques and/or algorithms for disease diagnosis prediction from the perspective of cardiovascular disease (CVD)”, International Journal of Computing Algorithm, Volume 11, Issue 2, PP. 6-16, Dec. 2022, https://doi.org/10.20894/IJCOA.101.011.002.002.

[9] Gu, J, and, Oelke D, “Understanding Bias in Machine Learning”, ArXiv abs/1909.01866, 1st Workshop on Visualization for AI Explainability in 2018 IEEE.

[10] Sun W, Nasraoui O, Shafto P, “Evolution and impact of bias in human and machine learning algorithm interaction”, PLOS ONE, Volume 15, Issue 8, Id - e0235502. https://doi.org/10.1371/journal.pone.0235502.

[11] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A, “A survey on bias and fairness in machine learning”, Volume 54, Issue 6, PP. 1-33, July 2021, https://doi.org/10.1145/3457607.

[12] Sun W, Nasraoui O, Shafto P, “Evolution and impact of bias in human and machine learning algorithm interaction”, PLOS ONE Volume 15, Issue 8, Id - e0235502. PP. 1-39, August 2020, https://doi.org/10.1371/journal.pone.0235502.

[13] Vokinger, KN, Feuerriegel S, & Kesselheim AS, “Mitigating bias in machine learning for medicine”. Communications Medicine 1, Volume 25, PP 1-3, August 2021 https://doi.org/10.1038/s43856-021-00028-w.

[14] Prajapati AK, & Singh, UK, “An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models. Journal of Systems Engineering and Information Technology (JOSEIT), Volume. 2, PP. 77-84. (2023) https://doi.org/10.29207/joseit.v2i2.5460.

[15] Mehrabi N, Morstatter F, Saxena N, Lerman K, and Galstyan A, “A Survey on Bias and Fairness in Machine Learning”, ACM Computing Surveys, Volume. 54, Issue 6, Article 115, PP 35, July 2022, https://doi.org/10.1145/3457607.

[16] Zhang K, Khosravi B, Vahdati S, Faghani S, Nugen F, Rassoulinejad-Mousavi SM, Moassefi M, M. Jagtap JM, Singh Y, Rouzrokh R, and J. Erickson B, "Mitigating Bias in Radiology Machine Learning: 2. Model Development", Radiology Artificial Intelligence, Volume 4, Issue 5, PP 1-8, August 2022, https://doi.org/10.1148/ryai.220010.

[17] Pagano TP, Loureiro RB, Lisboa FVN, Cruz GOR, Peixoto RM, De Sousa Guimarães GA, Dos Santos LL, Araujo MM, Cruz M, De Oliveira ELS, Winkler I, and Nascimento EGS, "Bias and unfairness in machine learning models: a systematic literature review", Volume 3, PP 1-24 Nov 2022.

[18] Gupta GK and Sharma DK, "A Review of Overfitting Solutions in Smart Depression Detection Models," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE - New Delhi, India, PP 145-151, March 2022, https://doi.org/10.23919/INDIACom54597.2022.9763147.

[19] Thomas H, Dignum V, and Bensch S. "Bias in Machine Learning What is it Good for?", Volume 2, PP 1-8, April 2020.

[20] Mehrabi N, Morstatter F, Saxena N, Lerman K, and Galstyan A,"A Survey on Bias and Fairness in Machine Learning", ACM Computing Surveys, Volume 54, Issue 6, Article No. 115, PP 1-35, July 2021, https://doi.org/10.1145/3457607.

[21] Fahse, T, Huber, V, Van Giffen B, "Managing Bias in Machine Learning Projects". Innovation Through Information Systems, Volume 2, Issue 2021, PP 94-109, Springer International Publishing, October 2021, https://doi.org/10.1007/978-3-030-86797-3_7.

[22] Bailly A, Blanc C, Francis E, Guillotin T, Jamal F, Wakim B, Roy P, “Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models”, Computer Methods and Programs in Biomedicine, Volume 213, PP 106504, ISSN 0169-2607, January 2022, https://doi.org/10.1016/j.cmpb.2021.106504.

[23] Gavrilov AD, Jordache A, Vasdani M, and Deng J, “Preventing Model Overfitting and Underfitting in Convolutional Neural Networks”, International Journal of Software Science and Computational Intelligence, Volume 10, Issue 4, PP 1-10, December 2018 https://doi.org/10.4018/IJSSCI.2018100102.

[24] Heintz F, Milano M, and O’Sullivan B “Trustworthy AI-Integrating Learning, Optimization, and Reasoning”, Conference proceedings, First International Workshop workshop, Springer Nature, PP 31-42, September 2020, https://doi.org/10.1007/978-3-030-73959-1.

[25] Li L, and Spratling M "Understanding and combating robust overfitting via input loss landscape analysis and regularization", Pattern Recognition, Volume 136, Issue, PP 1-11, April 2023, https://doi.org/10.1016/j.patcog.2022.109229.

[26] Gupta GK and Sharma DK, "A Review of Overfitting Solutions in Smart Depression Detection Models", 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022, PP 145-151, https://doi.org/10.23919/INDIACom54597.2022.9763147.

[27] Ghosh P, Azam S, Jonkman M, Karim S, Shamrat FMJM, Ignatius E, Sultana S, Beeravolu AR, and De Boer AF, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," IEEE Access, Volume. 9, PP 19304-19326, 2021, https://doi.org/10.1109/ACCESS.2021.3053759.

[28] Mat Deris A and Zain AM and Sallehuddin R, "Overview of Support Vector Machine in Modeling Machining Performances", International Conference on Advances in Engineering, Procedia Engineering 2011, Volume 24, PP 308-312, https://doi.org/10.1016/j.proeng.2011.11.2647.

[29] Li H. "Support Vector Machine, Machine" Learning Methods. Springer, Singapore, 2024 https://doi.org/10.1007/978-981-99-3917-6_7.

[30] Yang Xu, Bern Klein, Genzhuang Li, Bhushan Gopaluni, "Evaluation of logistic regression and support vector machine approaches for XRF based particle sorting for a copper ore", Minerals Engineering, Volume 192, PP 108003, ISSN 0892-6875, 2023, https://doi.org/10.1016/j.mineng.2023.108003.

[31] Loh, WY, "Logistic Regression Tree Analysis", Pham, H. (eds) Springer Handbook of Engineering Statistics, Springer Handbooks, Springer, London, 2023 https://doi.org/10.1007/978-1-4471-7503-2_30.

[32] Jain A, and Sharma A, "Membership function formulation methods for fuzzy logic systems: A comprehensive review" Journal of Critical Reviews, Volume 7, Issue 19 PP 8717-8733, 2020.

[33] Subhashini, LDCS Li, Y, Zhang J, and Atukorale, AS, "Integration of fuzzy logic and a convolutional neural network in three-way decision-making", Expert Systems with Applications, Volume 202, PP 117103, 2022, https://doi.org/10.1016/j.eswa.2022.117103.