Diabetes Prediction Medicament using Optimized SVM algorithm with Outlier detection and removal

K. Kanmani,
Research scholar, Dr. Ambedkar Govt Arts College, Vyasarpadi , Chennai- 600039, India,
Assistant professor, Department of Computer Applications, College of Science and Humanities, SRM Institute of Science and Technology, kattankulathur- 603203. Chennai, TN, India
Dr.A. Murugan
Associate Professor and Head,
PG and Research Department of Computer Science, Dr.Ambedkar Govt. ArtsCollege(Autonomous), Vyasarpadi,
Chennai – 600039, TN, India

Abstract
Data mining-based prediction techniques can help in the early detection of diabetes and the related critical events. Data mining algorithms are being used to improve the early diagnosis of diseases such as type 2 diabetes. This study aims to develop an Optimized Prediction Model that can predict the risk of diabetes. In this work we are using an optimized SVM with applied outlier detection removal mechanism based on the input factors from individuals, we implemented this process by Using R-Programming.

Keywords: *Data mining, SVM, Outlier Detection, R programming*

I. Introduction

Data mining-based forecasts for diabetes diagnosis and treatment can help in the early detection of the disease and its related critical events. [1, 14] This field of study has numerous applications in diabetes management and treatment. Data mining techniques for the early detection and predicting of diabetes can help minimize its complications and improve its management. To predict T2DM, diabetes complications, genetic background, health care, and management of T2DM, several data mining algorithms were applied. Other diseases, such as cancer and cardiovascular disorders, have employed similar strategies for prognosis and prediction. It is, however, always difficult to choose the best suited prediction approaches for a certain task. As a result, we chose ensemble-based methods that have recently been utilised in similar investigations and demonstrated the best outcomes, particularly in terms of outlier detection [3,4].

1.Analysis of Extreme Values

Extreme Value Analysis is the simplest basic kind of outlier detection and is ideal for data with only one dimension. Outliers are supposed to be values that are either too large or too little in this Outlier analysis approach. Classic examples have proven the analysis of extreme values [1]. include the Z-test and the student's t-test [5,6,7]. These are useful heuristics for preliminary data analysis, but they are less useful in multivariate situations. Extreme Value Analysis is frequently used to interpret the results of various outlier detection methods [1].

2.Models that are linear

The data is modelled into a lower-dimensional sub-space using linear correlations in this approach. Then, for each data point, the distance to a plane that suits the sub-space is determined [8]. Outliers are found using this distance. PCA (Principal Component Analysis) is a type of linear model used to discover anomalies [1].

3.Statistical and Probabilistic Models

Probabilistic and Statistical Models assume certain statistical parameters in this method. To estimate the parameters of the model, they use expectation-maximization (EM) strategies [9]. Finally, they compute the likelihood of each data point belonging to the determined distribution. Outliers are points with a low chance of belonging to a group [1].

4.Models Based on Proximity

Outliers are represented as separate points from the rest of the data in this manner. The most common methodologies of this type include cluster analysis, density-based analysis, and closest neighbour analysis [1,10,13].

5.Models Based on Information

Outliers raise the minimum code length required to describe a data set in this manner [1].

In data mining, finding outliers is a critical task. Outlier detection, as a subcategory of data mining, provides a wide range of applications and requires greater attention from the data mining community [14]. Data mining is the process of extracting high-quality, valuable information from unstructured data using algorithms from data mining, machine learning, statistics, and natural language

Manuscript received May 5, 2022
Manuscript revised May 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.5.74

processing. The use of text mining for corporate applications must have risen significantly in recent years. The reason behind this is because more people are becoming aware of text mining and the cheaper costs at which text mining machines have become available [11].

II. Related Work

Regarding application of DBSCAN for outlier detection, several studies have been conducted and showed significant results in identifying outliers as well as improving the classification result. Past literature showed that by removing noise the quality of real datasets is enhanced [13]. Support vector data description (SVDD) was utilized to classify the dataset. The University of California, Irvine (UCI) dataset has been utilized for the experimental scenario and the proposed method showed an efficient result. In the case of social network, ElBarawy et al. utilized DBSCAN to emphasize community detection. The result showed that the DBSCAN successfully identifies outliers [14]. Eliminating the outliers prompts a precise clustering result that assists with the community identification issue in the area of social network analysis. The DBSCAN-based outlier detection also showed significant results on detecting the outlier sensor data. Alfian et al. proposed a real-time monitoring system that is based on smartphone sensors for perishable food [15]. As outliers arise in sensor data due to inadequacies in sensing devices and network communication glitches, Alfian et al. used outlier detection that is based on DBSCAN to refine the outlier data. The findings demonstrated that DBSCAN was utilized to effectively recognize/characterize outlier data as isolated from normal sensor data. Abid et al. proposed outlier detection based on DBSCAN for sensor data in wireless sensor networks [16]. The proposed model successfully separated outliers from normal sensors data. Based on the experiment on synthetic datasets, their proposed model showed significant results in detecting outliers, obtained an accuracy. Finally, Tian et al. proposed an outlier detection method of soft sensor modeling of time series [17]. To remove inaccurate clustered data, the modified K-means method was utilized, followed by the logistic regression technique.

III. Methodology

An SVM training algorithm creates a model that assigns new examples to one of two categories, making it a non-probabilistic binary linear classifier, given a series of training examples that are individually designated as belonging to one of two categories. Figure1, Shows that pre-processed data without outliers. The goal of using SVMs is to discover the correct line in two aspects or the best hyperplane in more than two aspects to assist us in sorting our space. The maximum margin and the maximum distance between data points of both classes, is used to find the hyperplane (line). We used data pre-processing to remove missing values and pick the most significant feature that contributes to model accuracy. Data mining contributes a requisite role for identifying ineffective attribute on the clinical data set, which could provide valuable knowledge base for efficient and successful decision making [12].

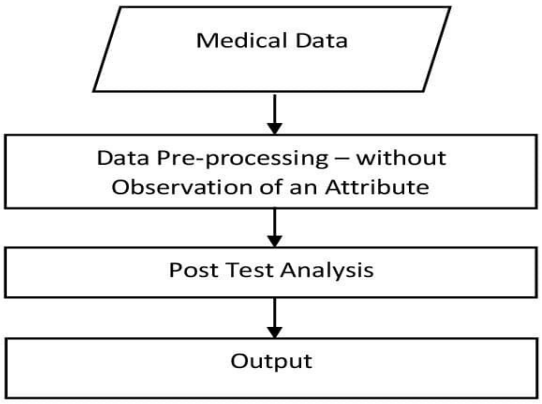

Figure 1: Preprocessing - without outliers

Patients typically visit a diagnostic center, consult with their doctor, then wait a day or more for their results. Furthermore, they must squander their money in vain every time they need to obtain their final report [18]. However, with the advancement of Data Mining techniques, we may be able to find a solution to this problem; we have developed a framework based on data mining that may predict if a patient has diabetes or not.[19].

SVM refers for Support Vector Machine and is one of the most common supervised machine learning models used in classification. A support vector machine's goal is to determine the optimal highest-margin separation hyperplane between two classes given a two-class training sample [20]. Hyperplane should not be closer to data points belonging to the other class for better generalization. A hyperplane that is far from the data points in each category should be chosen. The support vectors are the points that are closest to the classifier's margin [21]. The experiment's accuracy is assessed using R – Programming. Sample Pseudo code for this work is shown in Figure 3. Classifying an attribute based on the age and Glucose Level, determined that significant attributes only have a major role for the prognostication of the diabetes.

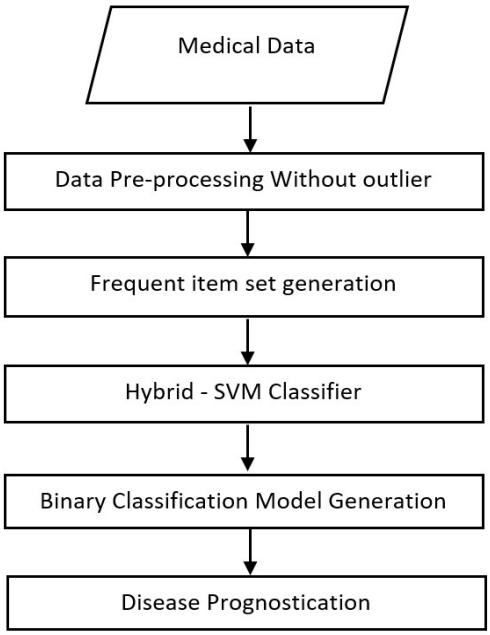

Figure 2 : Proposed Model - disease prognostication

The proposed method uses Support Vector Machine (SVM), with outlier removal as the classifier for diagnosis of diabetes shown in Figure 2. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease.

With the increasing prevalence of diseases such as diabetes and high blood pressure, Mining algorithms have been proposed to improve the early detection of these conditions [22]. The goal of this study is to develop an Optimised Predictive Model that can predict the type 2 diabetes and high blood pressure based on the input factors.

Data pre-processing is important since it can improve the accuracy of the classifier. Feature selection is used to pick a subset of features that make a significant contribution to the objective class. The goal of the feature selection technique is to improve accuracy while also reducing process length and cost computation [23,24]. Choosing relevant features from a dataset based on their estimated significance is one approach of doing so. Finally, the dataset can be cleaned up by removing unrelated features.

To boost performance, a variety of data preparation techniques is used. To avoid classification bias towards the typical cases, we first oversampled (with replacement) certain data occurrences (outliers) regardless of class. The OSVM algorithm was able to access the particular attributes embedded in outlier instances as a result of this. The increase in classification accuracy as a result of the data preprocessing combination. In compared to prior research that employed the same dataset in the literature, the proposed approach improves accuracy.

```
Begin
Load dataset values
    Training = = MyData[intrain,]
    testing = = MyData[-intrain,]
Examine training data and testing data
    MyData = =Null Removal (MyData)
    (Glucose, BP, BMI, Insulin, Age)
Removal of null values in specific attributes
Mydata = = Remove Insignificant data
Removal of insignificant data
    Check If Glucose Level < 95 assign the value 0
    Else If Glucose Level >=95 and Glucose Level<=140 assign the value 1
    Else IF Glucose Level >140 assign the value 2
If the value is 0 then the patient doesn't require diabetes treatment
Check all the attributes and Implement OSVM.
    Apply traincontrol(), and repeatedcv method , get an accuracy for OSVM
    Obtained confusion matrix result
    Print accuracy
End
```

Figure 3: Sample Pseudo code for outlier removal

Pima Indians Diabetes Database to predict whether or not a patient has diabetes dataset. The following are the parameters which involved in diabetes prediction [15]

1.Pregnancies
2.Glucose
3.BloodPressure
4.SkinThickness
5.Insulin
6. BMI
7. Diabetes pedigree Function
8. Age

These are the important attributes which is used to prognosticate the patient's diabetes level.

Data mining classifiers continue to struggle with learning from outliers and unbalanced data. Data preparation solutions are known to be efficient and simple to implement among the several techniques dedicated to addressing this problem. In this research, we present a selective data preprocessing strategy for achieving an equal distribution by embedding knowledge of outlier cases into an artificially produced subset. By inserting artificial minority occurrences, the Optimised SVM was utilised to balance the training data. But not until the outliers had been identified and

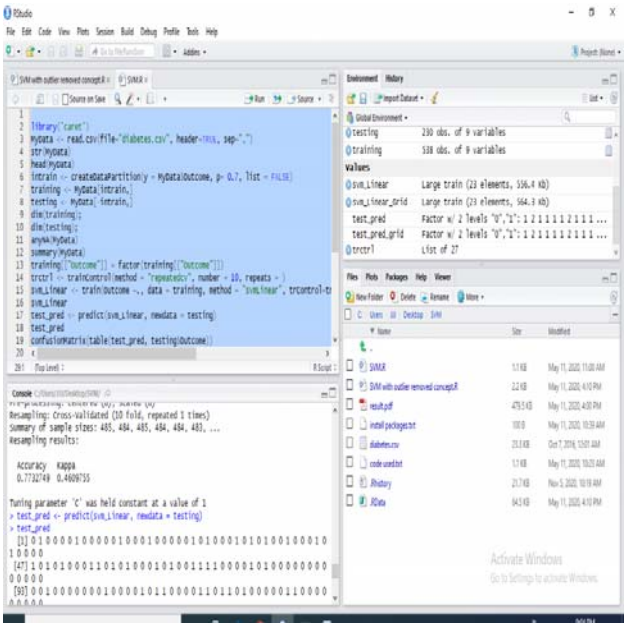

Figure 4: Without Mutation and outlier Detection

oversampled (irrespective of class). The goal is to keep the training dataset balanced while maintaining control

Implementation of SVM without outlier detection and Mutation which provides lower accuracy shown in Figure 4.

The learning process begins with the collection of data from various sources using multiple techniques. The next action is to prepare the data, or pre-process it, in order to resolve data-related difficulties and lower the space's dimensionality by deleting irrelevant data (or selecting the data of interest). Also, because amount of data utilised for learning is so large, it's difficult for the system to make decisions, thus algorithms are built to analyse the data and recover information from previous experiences using logic, probability, statistics, and control theory, among other things. The model is then tested to determine the system's correctness and performance

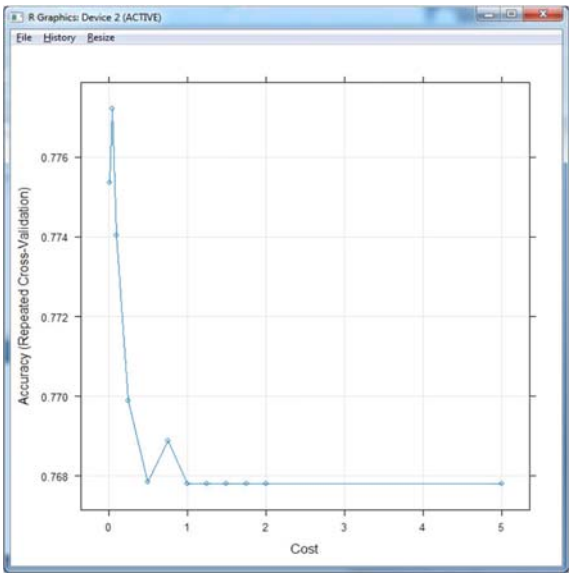

Figure 5: Accuracy Repeated cross validation

Outliers are being dealt with in a new way. Two optimization ways to deal with outliers in order to increase the dataset's quality, develop a highly accurate healthcare model, and save human lives. Detecting outliers at the dataset level for various classifications. Figure 5, shows the accuracy cross validation. The anomalies values are detected by computing the distribution of the full dataset and then removing the outliers' values based on that. However, because the values of different classes of the dataset have different outliers' calculations, this may present an issue.

Pattern mining techniques known as sequential pattern mining (SPM) are commonly used for the diagnosis of diabetes. They are based on the data collected during the course of study. [25] This Work can be utilized by healthcare professionals to make informed decisions about the insulin treatment plan. It can be used by the users to generate a customized set of sequences based on their blood glucose levels

Prior to processing data, there was a certain assumption that was used to set the upper and lower boundary sets. It is generally assumed that random oversampling is applied to data that are imputed. The generic assumption that we have considered when processing data is that the data is real and that its approach removal of anomaly data.

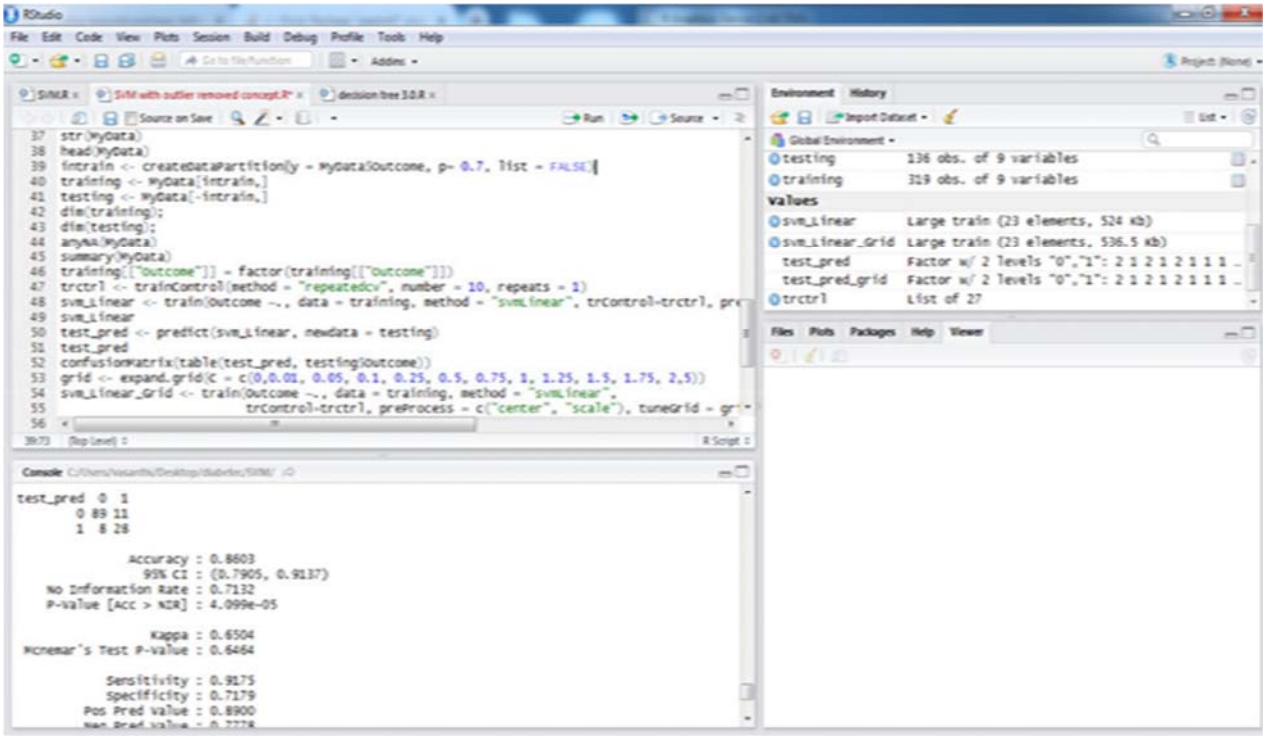

Figure 6: Proposed work –OSVM with Outlier Detection

IV. Conclusion

In this paper, Applied OSVM method that aims to remove the outliers by extracting knowledge from the data pre-processing stage. Learning from imbalanced and anomalous data is a major challenge for classification systems. The first phase of this study focused on identifying the best way to handle missing data. The second phase analyzed the effects of different data imputation methods on classification model performance. This paper proposes a method that combines Optimal support vector machine (OSVM) and Outlier detection and removal techniques to improve the accuracy of classification methods. It does so by mapping the features of the data with high accuracy

For more effective data classification, an outlier detection method is utilized to remove misclassified instances, Data mining algorithm using an Optimized SVM model. Figure 6, shows an accuracy value of our work proposed work, 86%. It is far better result compare to the previous SVM (without outlier Detection) algorithm which has only 78%. Outlier Detection has made it's a way into the scientific world and it is revolutionizing how to operate. Its ability to collect and analyses large amounts of data has a wide range of applications in healthcare and other fields. With the PIMA diabetic dataset, the results suggest that an Optimized SVM performs better. The anomaly prediction can be performed from the input data, which cannot have anomaly. It can then be used to warn the patient and the doctor about the seriousness of the situation, so the patient can elevate early medicament and gets better treatment.

References

[1] http://www.digitalvidya.com
[2] http://www.healthdata.org/sites/default/files/files/2017_India_StateLevel_Disease_Burden_Initiative_Full_Report%5B1%5D.pdf
[3] The increasing burden of diabetes and variations among the states of India: The Global Burden of Disease Study 1990–2016. https://doi.org/10.1016/S2214-109X(18)303875
[4] Rahman, R. M., Afroz, F. (2013), "Comparison of various classification techniques using different data mining tools for diabetes diagnosis". *Journal of Software Engineering and Applications* 6 (03): 85.
[5] Kumari, V, Chitra, R. "Classification of diabetes disease using support vector machine", *International Journal of Engineering Research and Applications* 2013;3(2):1797-1801.
[6] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on

data mining", *Informatics in Medicine Unlocked*, 2018, Pages 100-107.

[7] Sneha, N., Gangil, T. "Analysis of diabetes mellitus for early prediction using optimal features selection". *J Big Data 6*, 13 (2019). https://doi.org/10.1186/s40537-019-0175-6

[8] Dutta, Debadri and Paul, Debpriyo and Ghosh, Parthajeet. "Analysing Feature Importances for Diabetes Prediction using Machine Learning". 924-928. 10.1109/*IEMCON*.2018.8614871, 2018.

[9] Sharma, Himani and Kumar, Sunil. "A Survey on Decision Tree Algorithms of Classification in Data Mining". *International Journal of Science and Research* (IJSR). 5, 2016.

[10] Ijaz MF, Alfian G, Syafrudin M, "Rhee J. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest". *Applied Sciences*. 2018; 8(8):1325. https://doi.org/10.3390/app8081325

[11] Maniruzzaman, M., Rahman, M.J., Al-MehediHasan, M. *et al*. "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers." *J Med Syst* **42,** 92 (2018). https://doi.org/10.1007/s10916-018-0940-7

[12] K. Kanmani, Dr. A. Murugan, "Prognosticate and diagnosis of diabetes using data preprocessing and null value removal on the modified data set with possible outcome of a decision tree construction through R- programming", *Materials Today: Proceedings*,2020, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.10.631.

[13] "IDF Diabetes Atlas" - 8th Edition, International Diabetes Federation, 2017. [Online]. Available: *https://diabetesatlas.org/*. [Accessed: 15- Dec2018].

[14] Global Report on Diabetes Who Library Cataloguing-in-Publication Data Global report on diabetes. 2016.

[15] "PIMA Indian Diabetes Dataset, An open dataset," *UCI Machine Learning Repository*. [Online]. Available: http://ftp.ics.uci.edu/pub/machine

[16] Nonso Nnamoko, Ioannis Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction", *Artificial Intelligence in Medicine*, 104,2020

[17] Roopesh Padmaraju Alluri, Dr. Hemavathy,"Diabetes Prediction Using Ensemble Techniques", *International Journal of Applied Engineering Research* ISSN 0973-4562,16, Number 5 (2021) pp. 410-415

[18] M. Rout and A. Kaur, "Prediction of Diabetes Risk based on Machine Learning Techniques," *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, pp. 246-251, doi: 10.1109/ICIEM48762.2020.9160276.

[19] Bhatia, Kanika & Syal, Rupali, "Predictive analysis using hybrid clustering in diabetes diagnosis", (2017) 447-452. 10.1109/*RDCAPE*.2017.8358313.

[20] Jahangir, M and Afzal, H and Ahmed, M and Khurshid, K and Nawaz, R (2018) "An expert system for diabetes prediction using auto tuned multi-layer perceptron". *IEEE Intelligent Systems*. pp. 722-728. ISSN 1541-1672

[21] Zhou, H., Myrzashova, R. and Zheng, R. "Diabetes prediction model based on an enhanced deep neural network". *J Wireless Com Network* 2020, 148 (2020). https://doi.org/10.1186/s13638-020-01765-7

[22] Samir Kendale, Prathamesh Kulkarni, Andrew D. Rosenberg, Jing Wang; "Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension". *Anesthesiology* 2018; 129:675–688 doi: https://doi.org/10.1097/ALN.0000000000002374

.[23] Muhammad, L.J., Algehyne, E.A. and Usman, S.S. "Predictive Supervised Machine Learning Models for Diabetes Mellitus". *SN COMPUT. SCI.* 1**,** 240 (2020). https://doi.org/10.1007/s42979-020-00250

[24] Rajagopalan A, Vollmer M. "Rapid detection of heart rate frag-mentation and cardiac arrhythmias: cycle-by-cycle or analysis, supervised machine learning model and novel insights. In: RiañoD, Wilk S, ten Teije A, editors. Artificial intelligence in medicine". AIME 2019. *Lecture notes in computer science. Springer, Cham.2019*. p. 11526.

[25] Ye Q, Qin L, Forgues M, et al. "Predicting hepatitis B virus–positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning". *Nat Med*. 2003; 9:416