Journal of Big Data

RESEARCH Open Access

Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms

Ghada Mostafa[1,3*], Hamdi Mahmoud[1*], Tarek Abd El-Hafeez[2,3*] and Mohamed E. ElAraby[1*]

*Correspondence:
GhadaMostafaAbdelaziz8_pg@fcis.bsu.edu.eg; Dr_hamdimahmoud@yahoo.com; tarek@mu.edu.eg; Mohamed.elaraby@fcis.bsu.edu.eg

[1] Computer Science Department, Faculty of Computers and Artificial Intelligence, Beni-Suef National University, Beni-Suef, Egypt
[2] Department of Computer Science, Faculty of Science, Minia University, El-Minia, Egypt
[3] Computer Science Unit, Deraya University, El-Minia, Egypt

Abstract

Hepatocellular carcinoma (HCC) is a highly prevalent form of liver cancer that necessitates accurate prediction models for early diagnosis and effective treatment. Machine learning algorithms have demonstrated promising results in various medical domains, including cancer prediction. In this study, we propose a comprehensive approach for HCC prediction by comparing the performance of different machine learning algorithms before and after applying feature reduction methods. We employ popular feature reduction techniques, such as weighting features, hidden features correlation, feature selection, and optimized selection, to extract a reduced feature subset that captures the most relevant information related to HCC. Subsequently, we apply multiple algorithms, including Naive Bayes, support vector machines (SVM), Neural Networks, Decision Tree, and K nearest neighbors (KNN), to both the original high-dimensional dataset and the reduced feature set. By comparing the predictive accuracy, precision, F Score, recall, and execution time of each algorithm, we assess the effectiveness of feature reduction in enhancing the performance of HCC prediction models. Our experimental results, obtained using a comprehensive dataset comprising clinical features of HCC patients, demonstrate that feature reduction significantly improves the performance of all examined algorithms. Notably, the reduced feature set consistently outperforms the original high-dimensional dataset in terms of prediction accuracy and execution time. After applying feature reduction techniques, the employed algorithms, namely decision trees, Naive Bayes, KNN, neural networks, and SVM achieved accuracies of 96%, 97.33%, 94.67%, 96%, and 96.00%, respectively.

Keywords: Deep learning, Machine learning, Hepatocellular carcinoma, Liver cancer, Feature selection, Artificial Intelligence

Introduction

According to reports from the World Health Organization (WHO), approximately 14.1 million individuals are diagnosed with cancer each year, resulting in 8.2 million deaths globally [1]. Hepatocellular carcinoma (HCC) is a form of liver cancer that arises from chronic liver disease and cirrhosis. Recent studies indicate that HCC is the most lethal cancer worldwide, leading to approximately 600,000 deaths annually [2]. Furthermore, liver cancer holds the sixth position among the most frequently diagnosed cancers worldwide

[3]. These facts demonstrate the global impact of HCC on human lives. Consequently, it is crucial to reduce the mortality rate associated with HCC, which can only be achieved through early detection. To accomplish this goal, it is imperative to leverage various data mining and machine learning techniques to develop an automated diagnostic system that can accurately predict HCC, ensuring more efficient and timely detection. Data mining is a multidisciplinary domain that employs principles from computer science and statistics to extract valuable information, such as features or rules, from provided data [4]. Conversely, machine learning is a branch of computer science that focuses on techniques and methodologies through which machines acquire knowledge and learn from experience [5]. In the present era, machine learning techniques and data mining are experiencing rapid growth and extensive application in the realm of medical diagnostics to tackle various challenges such as [6–13].

Our research began with a focus on acknowledging the importance of normalized data. A clear trend was observed in previous work—better model performance with normalized data. This observation led us to adapt our dataset accordingly. Next, we introduced feature selection methods, starting with the powerful "Recursive Feature Elimination (RFE)". This method tests the model's performance with each potential feature, systematically removing features and re-testing the model to find the best iteration. Next, we used "Principal Component Analysis (PCA)", which is a popular method for feature extraction. Its goal is to reduce the dimensionality of a data set while preserving as much of the information as possible. PCA accomplishes this by creating new uncorrelated variables or components that successively maximize variance. In our study, PCA was utilized to transform the data set into a set of linearly uncorrelated variables termed principal components. Finally, optimization feature operators were applied. It is well recognized that optimizing the selection of feature subsets can significantly improve the performance of a classifier. To rate the importance of a feature for the classification task, mutual information was utilized. This was followed by executing various machine learning algorithms to assess classification performance.

A clear challenge exists in the form of Hepatocellular Carcinoma (HCC)—a lethal form of cancer cloaked in diagnostic complexity. Accurate, efficient predictive models are crucial for timely diagnosis and optimized treatment. However, conventional predictive models are hindered by the 'dimensionality curse', a common obstacle in high-dimensional datasets used in HCC diagnosis.

Problem statement

Despite being one of the most lethal forms of cancer, Hepatocellular Carcinoma (HCC) remains shrouded in an air of diagnostic complexity. The development of accurate and efficient predictive models represents a critical facilitator of timely diagnosis and effective treatment. Stunted by the dimensionality curse commonly associated with high-dimensional datasets acquired in HCC diagnosis, traditional predictive models have demonstrated limited proficiency.

Research question

Can the application of alternative feature reduction techniques significantly enhance the performance of machine learning algorithms in the prediction of Hepatocellular Carcinoma?

Research gap

Previous studies have noted the positive relationship between reducing feature dimensionality and the predictive accuracy of machine learning algorithms. However, there remains a conspicuous lack of comprehensive approaches that compare the performance of various machine learning algorithms under the influence of different feature reduction techniques in the domain of hepatocellular carcinoma prediction.

Contributions

This study heralds an important contribution to the field of computational HCC prediction by comparing the performance of much-utilized machine learning algorithms before and after the implementation of feature reduction techniques. The main contributions can be summarized as follows:

1. Adoption of data normalization to improve our model's performance, as reinforced by earlier studies.
2. Execution of feature selection methods including 'Recursive Feature Elimination (RFE)' and 'Principal Component Analysis (PCA)' to boost the effectiveness of our predictive model.
3. Assessment of the influence of various features on the task of classification by deploying mutual information.
4. Conducting a performance comparison of differing machine learning algorithms, gauging their classification results.
5. Addressing existing research shortcomings by performing an extensive comparison of multiple feature reduction techniques and their corresponding impact on the outcomes of a range of machine learning algorithms, particularly about Hepatocellular Carcinoma (HCC) prediction.
6. Advancing the computational prediction field for HCC by examining performance shifts in a variety of machine learning algorithms both before and after the integration of feature reduction techniques.

Related work

In a research study by Abajian et al. [14] a study involving 36 patients with HCC who underwent transarterial chemoembolization. They employed machine learning techniques, specifically linear regression, and random forest, and achieved an overall accuracy of 78%. In a study by Ioannou et al. [15] focused on predicting the occurrence of hepatocellular carcinoma (HCC) within 3 years, a recurrent neural network (RNN) was trained using data from patients with hepatitis C virus (HCV)-related cirrhosis. The dataset included four variables measured at the beginning of the study and 27 variables measured over time, collected from 48,151 patients receiving healthcare within the US Department of Veterans Affairs system. The findings of the study demonstrated that the RNN model outperformed logistic regression in predicting the development of HCC within the specified timeframe. The RNN achieved an accuracy of 75.9% for all patients and 80.6% for patients who achieved sustained virologic response (SVR) in predicting the onset of hepatocellular carcinoma (HCC).

In a research study conducted by Nam et al. [16], a deep neural network was developed to predict the occurrence of hepatocellular carcinoma (HCC) over a 3- and 5-year period in patients with hepatitis B virus (HBV)-related cirrhosis who were undergoing entecavir therapy. The study examined 424 patients and demonstrated that the deep learning (DL) model outperformed six other previously reported models that utilized older modeling techniques. Additionally, the DL model was tested on a validation cohort consisting of 316 patients, and the results indicated a Harrell's C-index of 0.782, indicating a high level of accuracy in predicting the incidence of HCC in these patients.

Nam et al. [17] built upon their previous work by developing MoRAL-AI, a novel artificial intelligence model utilizing deep learning techniques, to identify liver cancer (HCC) patients at high risk of tumor recurrence after transplantation. The MoRAL-AI model analyzed several prognostic factors including tumor size, patient age, blood alpha-fetoprotein (AFP) levels, and prothrombin time to generate risk predictions. Results of the study demonstrated that MoRAL-AI outperformed traditional prediction models such as the Milan, UCSF, up-to-seven, and Kyoto criteria in determining which HCC patients faced elevated recurrence risk post-transplant. Specifically, MoRAL-AI achieved a C-index of 0.75 for prognostic accuracy compared to 0.64, 0.62, 0.50, and 0.50 for the other models respectively, with this difference being statistically significant ($p < 0.001$). In summary, MoRAL-AI represented an improved approach for identifying HCC patients likely to experience recurrence following liver transplantation.

In their study, Ali et al. [18] evaluated the predictive performance of various machine learning algorithms for hepatocellular carcinoma (HCC), including logistic regression, k-nearest neighbors (KNN), decision tree, random forest, and support vector machine (SVM). Additionally, they proposed and tested a novel combination approach utilizing linear discriminant analysis (LDA), genetic algorithm (GA), and SVM. When comparing all models, the results demonstrated the LDA-GA-SVM approach yielded the best overall predictive ability. Specifically, the LDA-GA-SVM achieved the highest accuracy of 0.899, sensitivity of 0.892, and specificity of 0.906. These performance metrics were superior to those obtained when using the other individual algorithms evaluated—logistic regression, KNN, decision tree, random forest, and SVM alone. Therefore, the study findings suggested the LDA-GA-SVM composite model may be the most effective machine learning-based predictive tool for HCC compared to the alternative algorithms analyzed.

Cao et al. [19] evaluated the predictive performance of various machine learning models—logistic regression, k-nearest neighbors (KNN), decision tree (DT), naïve Bayes (NB), and deep neural network (DNN)—using the original dataset. The accuracy of the models ranged from 57.5 to 70.6%. Precision varied between 40.7 and 70.1%, while recall rates were between 20.0 and 67.7%. False positive rates fell between 10.7 and 35.0% and standard deviation values ranged from 0.026 to 0.058. Among the models trained on the original dataset, KNN exhibited the best overall predictive ability. Specifically, KNN achieved an accuracy of 70.6%, precision of 70.1%, recall rate of 51.9%, and a false positive rate of 16.0% with a standard deviation of 0.042. These results indicate that of the algorithms tested on the unmodified data, KNN provided the most accurate and reliable predictions of disease status.

In a study by Zhang et al. [20] 237 patients with liver cancer, almost 39% (92 patients) were identified as having a positive marker for MVI. This group, with an average age of 52, was predominantly male (86 out of 92). The remaining 61% of patients (145 patients) were MVI-negative, with an average age of 54 and a more balanced male-to-female ratio (124 males to 21 females). Patients with MVI had larger tumors, a higher occurrence of tumor capsules, and elevated levels of certain proteins compared to those without MVI.

In a study by [21] After conducting machine learning analysis, they identified eight key feature variables (age, intratumoral arteries, alpha-fetoprotein, pre-operative blood glucose, number of tumors, glucose-to-lymphocyte ratio, liver cirrhosis, and pre-operative platelets) to develop six distinct prediction models. Among these models, the XGBoost model exhibited superior performance, as evidenced by the area under the receiver operating characteristic curve (AUC-ROC) values of 0.993 (95% confidence interval: 0.982–1.000), 0.734 (0.601–0.867), and 0.706 (0.585–0.827) in the training, validation, and test datasets, respectively. Furthermore, calibration curve analysis and decision curve analysis demonstrated that the XGBoost model exhibited favorable predictive performance and possessed practical value in clinical applications.

Motivated by the development of different diagnostic systems based on machine learning models to improve the precision of decision-making about HCC diagnosis and prediction we also conducted an approach to enhance hepatocellular carcinoma (HCC) prediction through Feature reduction methods. This study highlights the effectiveness of feature reduction in boosting the performance of various AI techniques for HCC nodule prediction. By streamlining the data, they were able to significantly improve the accuracy of algorithms like Naive Bayes, Neural Networks, Decision Tree, SVM, and KNN.

Materials and methods

Database description

Clinical patient data from the Cancer Genome Atlas (TCGA) database were used in this study, The TCGA LIHC clinical data set offers a robust resource for investigating the clinical landscape of hepatocellular carcinoma (HCC). This data, encompassing diverse patient demographics, tumor characteristics, treatment details, and clinical outcomes, facilitates a multi-faceted approach to understanding disease progression and informing research avenues [22–24].

- Patient demographics: Age, sex, ethnicity, socioeconomic factors, and medical history provide context for analyzing disease epidemiology and potential risk factors as shown in Fig. 1. Correlations between these variables and clinical outcomes can inform targeted prevention and early intervention strategies.
- Tumor characteristics: Detailed information on tumor size, stage, grade, location, and presence of underlying liver disease allows for stratification of patient populations and facilitates investigation of tumor progression patterns.

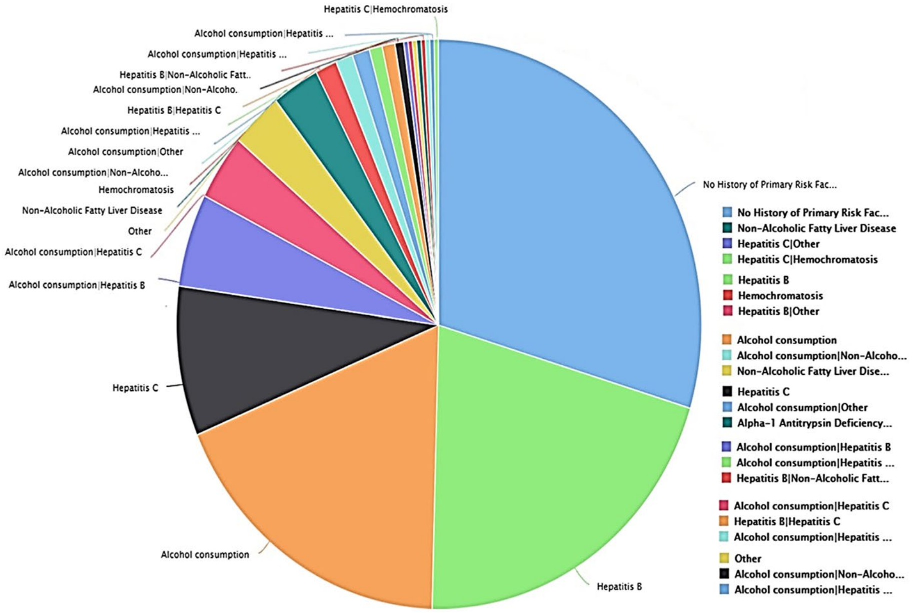


Fig. 1 Hepatocellular carcinoma risk factors history in TCGA LIHC data set

- Treatment details: Data on surgical procedures, radiation protocols, and chemotherapy regimens allows for comparative effectiveness studies and identification of optimal treatment strategies for different patient subgroups.
- Clinical outcomes: Data on overall survival, disease-free survival, time to recurrence, and response to treatment offer important endpoints for evaluating treatment efficacy and informing clinical decision-making.
- Limitations: While the TCGA LIHC clinical data set is comprehensive, it's important to acknowledge potential limitations due to data collection inconsistencies, missing follow-up data, and selection bias. Careful consideration of these limitations is necessary to ensure accurate interpretation of results and informed research conclusions.

The dataset employed in this study comprised 77 features for each of the 377 patients in total. The label of the dataset denotes tumor status and can assume a value of "tumor-free" or "with tumor". The term "tumor-free" does not imply a state of normalcy, but instead refers to the absence or persistence of the neoplasm (tumor). It represents a statement regarding the progression or lack thereof of the initial disease. It is crucial to mention that there are missing values for each feature in the dataset that have the information of all features. Within the existing body of literature, two distinct approaches are commonly employed to address missing values. The first method involves removing all samples that contain missing values, but this approach is not feasible in our case as it would result in the loss of a significant portion of the samples. Consequently, we opted to employ the imputation method to fill in the missing values. Missing data was addressed through a diverse range of imputation methods during the studies [25–28]. We utilized a statistical approach to impute missing values by substituting them with the mean value of the corresponding column or feature in which the missing value was found. Elaborate information is provided about the clinical features of the TCGA dataset in Table 1.

Methodology

The proposed research entails a multi-pronged approach to enhance hepatocellular carcinoma (HCC) prediction through Feature reduction methods including feature importance, hidden feature correlation, and feature selection [29] using different algorithms. The initial phase involved a thorough review of existing literature on deep learning applications in risk assessment, diagnosis, prognosis, and therapy for HCC patients. Subsequently, a meticulous analysis of clinical variables was conducted. Deep learning and machine learning algorithms were then implemented for HCC prediction, incorporating various feature reduction techniques. The overarching objective is to demonstrably validate the superiority of employing alternative feature selection methods compared to using all features within the machine learning models for achieving accurate HCC prediction.

In this study, the workflow for training a dataset using feature weight, feature correlation, Normalization, and optimization operators in RapidMiner [30] involves a series of steps designed to enhance the model-building process.

First, the dataset was loaded into RapidMiner, and the relevant operators were added to the process. The weights operator allows assigning importance or significance to individual instances or attributes in the dataset. This was useful when certain instances or attributes carry more weight or relevance in the analysis.

Next, the correlation operator was applied to identify and measure the relationships between different attributes in the dataset. It helps in understanding which attributes are strongly correlated with the target variable or with each other. This information can guide feature selection and eliminate redundant or highly correlated attributes, reducing the dimensionality of the dataset.

After the correlation analysis, the normalization operator was utilized to scale and standardize the numerical attributes in the dataset. This step ensures that all attributes have similar ranges and distributions, preventing any single attribute from dominating the model training process due to differences in their scales. Normalization enhances the stability and convergence of various algorithms leading to improved model performance.

Following normalization, the optimization operator was employed to select the most relevant subset of features from the dataset. It uses optimization algorithms and statistical measures to evaluate the contribution of each attribute to the model's performance. By iteratively evaluating different feature subsets, the optimization operator identified the combination of attributes that maximizes the model's accuracy or other defined performance metrics. This step helped in reducing noise, improving model efficiency, and enhancing interpretability.

Once the optimized feature subset was determined, the dataset was divided into training and testing sets 301 examples for train and 75 examples for test using appropriate sampling techniques.in our case, we used "Stratified sampling" which involves creating random subsets while ensuring that the distribution of classes within those subsets remains consistent with the overall class distribution in the entire example set.

Table 1 Information about the features of the TCGA dataset clinical variables

Features	Description	Type	Values
Ablation embolization tx adjuvant	Ablation embolization tx adjuvant	Binominal	No (364), Yes (13)
Age at diagnosis	Age at initial pathologic diagnosis	Integer	Min (16), Max (90)
ajcc metastasis pathologic pm	Pathologic M	Nominal	M0 (272), MX (101), M (4)
ajcc nodes pathologic pn	Pathologic N	Nominal	N0 (257), NX (115), N1 (4)
ajcc pathologic tumor stage	Pathologic stage	Nominal	Stage I (175), Stage II (87), Stage IIIA (65), Stage IIIB (9), Stage IIIC (9), Stage III (3), Stage IV (2), Stage IVB (2), [discrepancy] (2),Stage IVA (1)
ajcc staging edition	System version	Nominal	7th 231,6th 119, 5th 23, 4th 4
ajcc tumor pathologic pt	Pathologic T	Nominal	T1 185, T2 93,T3 45,T3a 29 T4 13,T3b 7,T2a 1,T2b 1 TX 1, [discrepancy] 1
Alpha fetoprotien at procurement	Laboratory procedure alpha-fetoprotein outcome value	Integer	Min (1), Max (2035400)
Alpha fetoprotien norm range lower	Laboratory procedure alpha-fetoprotein outcome lower limit of normal value	Integer	Min (0), Max (6)
Alpha fetoprotien norm range upper	Laboratory procedure alpha-fetoprotein outcome upper limit of normal value	Integer	Min (6), Max (44)
bcr patient barcode	bcr patient barcode	Nominal	Ex: TCGA-2V-A95S
bcr patient uuid	bcr patient uuid	Nominal	Ex: 0004D251-3F70-4395-B175-C94C2F5B1B81
Bilirubin total	Laboratory procedure total bilirubin result specified the upper limit of the normal value	Real	Min (0.100), Max (19)
Bilirubin total norm range lower	Laboratory procedure total bilirubin result specified a lower limit of normal value	Real	Min (0), Max (1)
Bilirubin total norm range upper	Laboratory procedure total bilirubin results in upper limit normal value	Real	Min (0.200), Max (21)
Birthdays to	Days to birth	Integer	Min (− 32,120),Max (− 5862)
Child–pugh classification	Child–Pugh classification grade	Nominal	A (223), B (21), C (1)
Clinical M	Clinical M	Nominal	[Not applicable] 377
Clinical N	Clinical N	Nominal	[Not applicable] 377
Clinical stage	Clinical stage	Nominal	[Not applicable] 377
Clinical T	Clinical T	Nominal	
Creatinine level preresection	Hematology serum creatinine laboratory result value in mg dl	Real	Min (0.400),Max (124)
Creatinine norm range lower	Laboratory procedure creatinine results lower the limit of normal value	Real	Min (0), Max (62)
Creatinine norm range upper	Laboratory procedure creatinine results in the upper limit of normal value	Real	Min (0.900), Max (120)
Days to initial pathologic diagnosis	Days to initial pathologic diagnosis	Integer	0
Death days to	Days to death	Integer	Min (− 1), Max (3258)

Table 1 (continued)

Features	Description	Type	Values
Definitive surgical procedure	Specimen collection method name	Nominal	Lobectomy 145 Segmentectomy, Multiple 89 Segmentectomy, Single 88 Other (specify) 26 Extended Lobectomy 25 No 3 Total Hepatectomy with Transplant 1
Disease code	Disease code	Nominal	[Not available] 377
ECOG score	Eastern Cancer Oncology Group	Integer	Min (0), Max(4)
Ethnicity	Ethnicity	Nominal	NOT HISPANIC OR LATINO 340 HISPANIC OR LATINO 18 Other 17 [Not available] 2
Extranodal involvement	Extranodal involvement	Nominal	[Not applicable] 377
Family history cancer indicator	Relative family cancer history ind 3	Binominal	NO 263 YES 114
Family history cancer number of relatives	Cancer diagnosis first-degree relative number	Integer	Min (0), Max (9)
Form completion date	Form completion date	Date -Time	Ranged from (20-12-2010) to (9-7-2015)
Gender	Gender	Binominal	MALE 255 FEMALE 122
Height cm at diagnosis	Height	Integer	Min (64), Max (196)
Hepatic inflammation adj tissue	Adjacent hepatic tissue inflammation extent type	Nominal	None 257, Mild 101, Severe 19
Histologic diagnosis	Histological type	Nominal	Hepatocellular Carcinoma 367 Hepatocholangiocarcinoma (Mixed) 7 Fibrolamellar Carcinoma 3
History of hepato carcinoma risk factors	History hepato carcinoma risk factor	Nominal	Most (no history of primary risk factors 112 Hepatitis B 78 Alcohol consumption 69 Hepatitis C 32 Alcohol consumption|Hepatitis B 20 Alcohol consumption|Hepatitis C 14 Other 12 Non-Alcoholic Fatty Liver Disease 11)
History neoadjuvant treatment	History of neoadjuvant treatment	Binominal	No 375 Yes 2
History other malignancy	Prior dx	Binominal	No 340 Yes 37
icd 10	icd 10	Nominal	C22.0 377
icd o 3 histology	icd o 3 histology	Nominal	8170/3 360, 8180/3 7 8171/3 4, 8174/3 4 8173/3 1, 8310/3 1
icd o 3 site	icd o 3 site	Nominal	C22.0 377
Informed consent verified	Informed consent verified	Nominal	YES 377
Ishak fibrosis score	Liver fibrosis ishak score category	Nominal	0—No Fibrosis 76 6—Established Cirrhosis 72 1,2—Portal Fibrosis 31 3,4—Fibrous Speta 30 5—Nodular Formation and Incomplete Cirrhosis 9

Table 1 (continued)

Features	Description	Type	Values
Last contact days to	Days to the last follow-up	Integer	Max(3675)
New tumor event dx indicator	New tumor event after initial treatment	Nominal	NO 279 YES 98
Other hepato carcinoma risk factors	History hepato carcinoma risk factors other	Nominal	Most (No 345 Smoking 6 Tobacco use 6 Cirrhosis 2)
Patient id	Patient id	Nominal	EX: 4072
Pharmaceutical tx adjuvant	Postoperative rx tx	Binominal	NO 362 YES 15
Platelet count pre-resection	Lab procedure platelet results specified value	Integer	Min (4), Max (499,000)
Platelet norm range lower	Laboratory procedure platelet results in a lower limit of normal value	Integer	Min (0), Max (163,000)
Platelet norm range upper	Laboratory procedure platelet results in the upper limit of normal value	Integer	Min (6), Max (450,000)
Project code	Project code	Nominal	[Not available] 377
Prospective collection	Tissue prospective collection indicator	Binominal	NO 249 YES 128
Prothrom time INR norm range lower	Laboratory procedure international normalization ratio results lower limit of normal value	Real	Min (0), Max (11)
Prothrombin time INR at procurement	laboratory procedure prothrombin time result value	Real	Min (0.800), Max (36.400)
Prothrombin time norm range upper	Laboratory procedure international normalization ratio results upper limit of the normal value	Real	Min (1), Max (15)
Race	Race	Nominal	WHITE 187 ASIAN 161 BLACK OR AFRICAN AMERICAN 17 Other 10 AMERICAN INDIAN OR ALASKA NATIVE 2
Radiation treatment adjuvant	Radiation therapy	Binominal	NO 373 YES 4
Residual tumor	Residual tumor	Nominal	R0 332, RX 22 R1 17,R2 1
Retrospective collection	Tissue retrospective collection indicator	Binominal	YES 249 NO 128
Serum albumin norm range lower	Laboratory procedure albumin results in a lower limit of normal value	Real	Min (0.300), Max (3800)
Serum albumin norm range upper	Laboratory procedure albumin result upper limit of normal value	Real	Min (0.500), Max (5100)
Serum albumin preresection	laboratory procedure albumin result specified value	Real	Min (0.200), Max (5200)
Stage other	Stage other	Nominal	[Not available] 377
Surgical procedure other	Surgical procedure name other specific text	Binominal	No 351 R hepatic lobectomy w/resection of L segment 1
Tissue source site	Tissue source site	Nominal	Most (DD 151)

Table 1 (continued)

Features	Description	Type	Values
Tumor grade	Neoplasm histologic grade	Nominal	G2 183,G3 124 G1 55,G4 13 [Not Available] 1
Tumor status	Person neoplasm cancer status	Binominal	TUMOR FREE 236 WITH TUMOR 141
Tumor tissue site	Tumor tissue site	Nominal	Liver 377
Vascular invasion	Vascular tumor cell invasion type	Nominal	None 230 Micro 94 Macro 17
Viral hepatitis serology	Viral hepatitis serology	Nominal	Most (no results 211)
Vital status	Vital status	Binominal	Alive 286 Dead 91
Weight kg at diagnosis	Weight	Integer	Min (40), Max (172)
Year of initial pathologic diagnosis	Year of initial pathologic diagnosis	Integer	Min (1995), Max (2013)

Finally, various modeling techniques, such as decision trees, Naive Bayes, KNN, neural networks, and SVM were applied to train the model using the selected features and the assigned weights.

Extracting meaningful insights from the TCGA LIHC dataset through regression tasks requires careful consideration of the chosen model. Several factors influence this selection, including data size, feature types, interpretability needs, and computational resources. For datasets with moderate sizes, similar to what might be encountered within TCGA LIHC, Naive Bayes offers a strong option. Decision trees are particularly well-suited for handling missing data inherent to real-world datasets, eliminating the need for extra imputation steps. K-Nearest Neighbors (KNN) stands out for its efficiency, directly comparing new data points to existing TCGA LIHC entries for prediction without a separate training phase. More complex models like neural networks can uncover hidden patterns within the data through automatic feature learning. Finally, Support Vector Machines (SVMs) offer robustness to noise, a common challenge in TCGA LIHC datasets. By carefully weighing these factors and evaluating model performance on the specific TCGA LIHC subset used, the model's performance is then evaluated using performance measures like accuracy, precision, F Score, and recall. A Summary of the Data Reduction Workflow for Predicting Hepatocellular Carcinoma, as Depicted in Fig. 2.

Results and discussion

Data preprocessing

The dataset initially consisted of 77 features. During the data cleaning process, 28 entries with unknown values in the "TUMOR status" column were replaced with "With TUMOR". In addition, two new features were introduced for further analysis: "optimal weight" based on Body Mass Index (BMI), categorized as Normal, Overweight, or Obesity, and "age stage" categorized as Middle Adulthood, Late Adulthood, or Young Adulthood. Redundant information such as age, height, weight, and other columns with repeated, unavailable, or inapplicable values, as well as patient IDs, were eliminated. As a result, the final dataset now comprises 59 features. Figure 3 illustrates the relationship

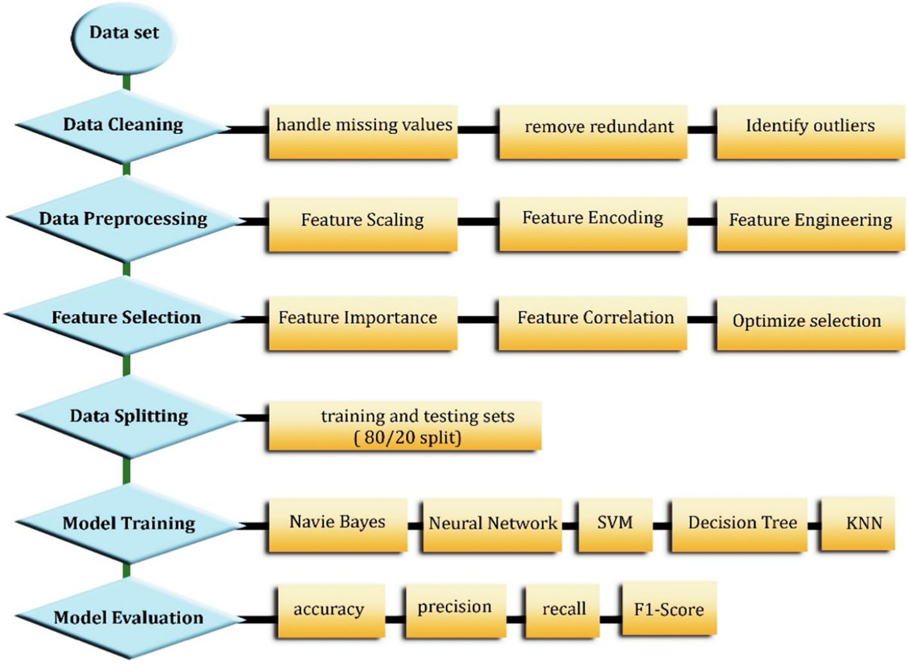


Fig. 2 Outline of data reduction workflow for Hepatocellular carcinoma Prediction

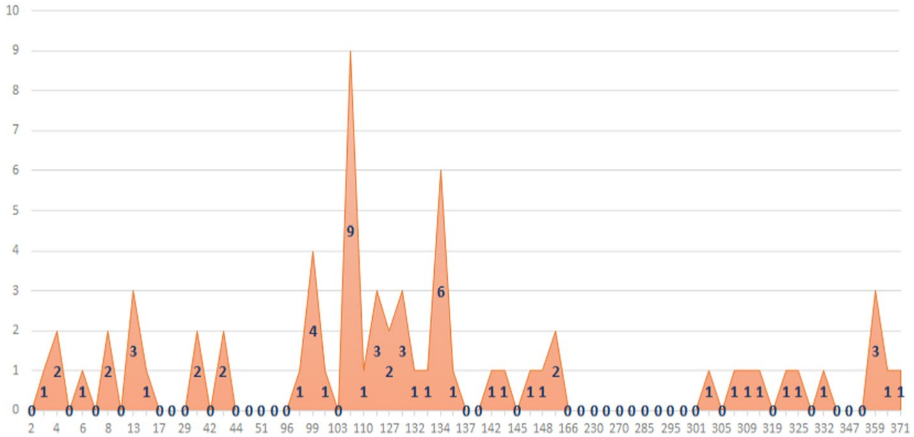


Fig. 3 Illustration of patients with obesity VS number of family relatives having a history of cancer

between patients with obesity and the number of family members with a history of cancer. Our findings indicate that the patient with obesity had the highest number of family members with this medical history.

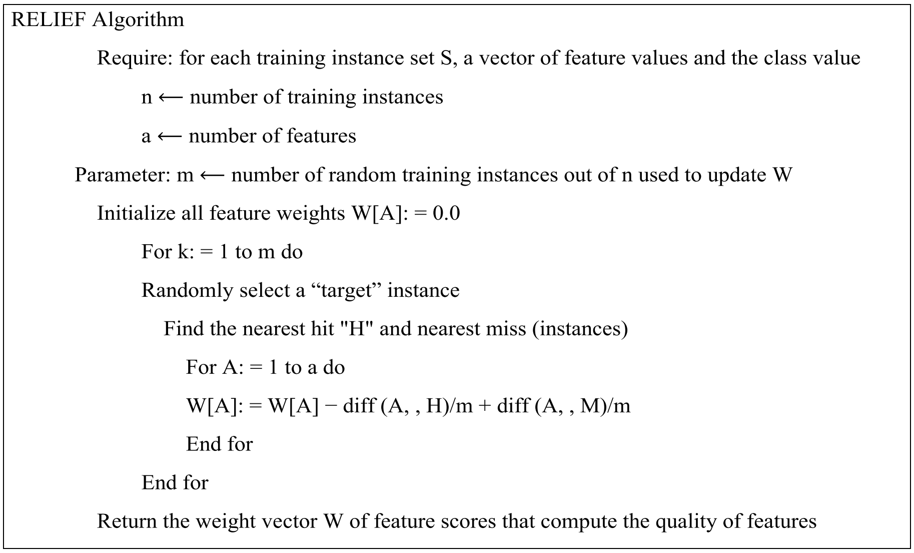
Feature importance

After data cleansing the remaining 59 features were weighted with different types of weight operators after replacing missing values using RapidMiner. First, we applied “Weight by Information Gain”. To determine how relevant each attribute is to the class attribute, the Weight by Information Gain operator uses a calculation called information gain [31]. Attributes with higher scores are considered more important.

While information gain is generally reliable for assessing attribute relevance [32], it does have a potential drawback. It can sometimes overestimate the importance of attributes that have a very large number of possible values. To overcome the limitations of information gain, particularly its sensitivity to attributes with numerous unique values, we used the information gain ratio by analyzing the information each attribute provides for understanding the target class, this method assigns weights that reflect their relative importance. The more insightful an attribute is for predicting the category, the higher its weight will be.

Secondly, we use the "Weight by Relief" operator. Considered one of the most effective and straightforward algorithms for evaluating feature quality, Relief has gained significant recognition. The fundamental concept behind Relief is to gauge the quality of features based on their ability to differentiate between instances of the same class and instances of different classes that are nearby[33, 34]. By sampling examples and comparing the feature values between the nearest examples of the same class and different classes, Relief calculates the relevance of features as described in [35].

Pseudocode of the Relief algorithm:



```
RELIEF Algorithm
        Require: for each training instance set S, a vector of feature values and the class value
            n ⟵ number of training instances
            a ⟵ number of features
      Parameter: m ⟵ number of random training instances out of n used to update W
        Initialize all feature weights W[A]: = 0.0
            For k: = 1 to m do
            Randomly select a “target” instance
              Find the nearest hit "H" and nearest miss (instances)
                For A: = 1 to a do
                W[A]: = W[A] − diff (A, , H)/m + diff (A, , M)/m
                End for
            End for
        Return the weight vector W of feature scores that compute the quality of features
```

Hidden feature correlation

Weight by Correlation is a feature selection methodology employed within the framework of Rapid Miner Studio [36]. This approach focuses on ascertaining the salience of features by quantifying their correlation with the target variable [37]. By assigning weights to individual features as shown in Fig. 4 based on their correlation coefficients, "Weight by Correlation" prioritizes those features that exhibit stronger

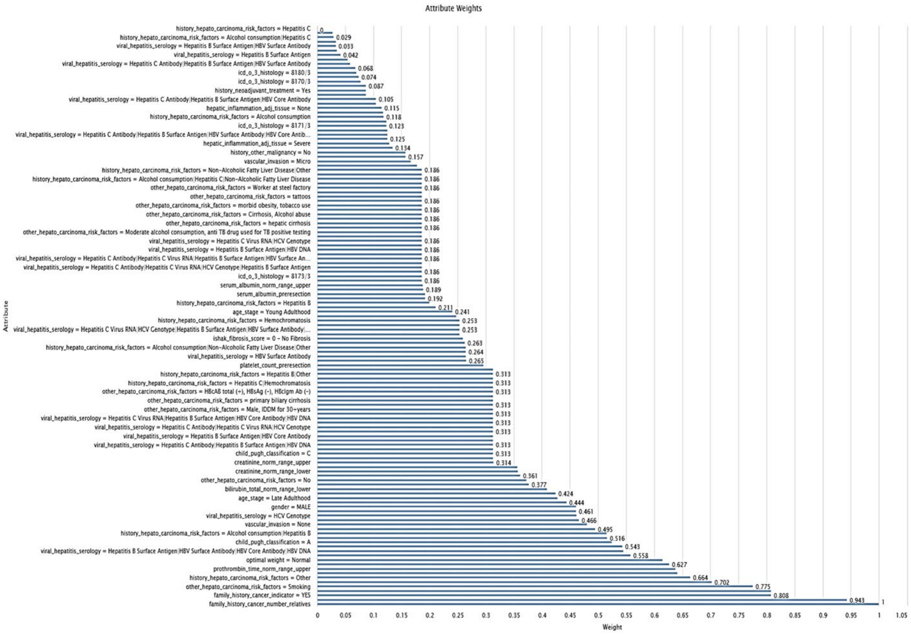


Fig. 4 Illustration of assigning weights to individual features based on their correlation coefficients

correlations. This weighting mechanism [38] facilitates the identification and selection of the most influential features, thereby enhancing the efficacy and precision of data analysis and modeling processes within Rapid Miner Studio.

Feature selection

Normalization is a technique employed to rescale values to fit within a specific range. It is particularly crucial when handling attributes that possess varying units and scales [39, 40].

The significance of data normalization in developing precise predictive models has been investigated across multiple machine learning algorithms [41], including Nearest Neighbors (NN) [42], Artificial Neural Networks (ANN) [43] and Support Vector Machines (SVM) [44]. Several researchers have confirmed the positive impact of data normalization on enhancing classification performance in various domains [45]. Examples include medical data classification [46, 47], multimodal biometrics systems [48], vehicle classification [49], faulty motor detection[50], stock market prediction [51], leaf classification [52], credit approval data classification [53], genomics [54], and other application areas [55, 56]. The purpose of the normalization operator is to perform the normalization process on selected attributes. There are four available normalization methods, with the "Range transformation" method being utilized in this case. This method normalizes all attribute values to a specified range [57]. Upon selecting this method, two additional parameters, namely "min" and "max," become visible in the parameters panel. The largest value in the attribute set is assigned to “max,” while the smallest value is assigned to “min.” All other values are proportionally scaled to fit within the provided range. It is worth noting that this method may be affected by outliers, as the boundaries adjust towards them. However, it retains the original distribution of the data points, making it suitable for data anonymization purposes as well.

Optimized selection is a valuable technique utilized in RapidMiner. This approach plays an essential role in streamlining the model-building process by automatically identifying and selecting the most relevant subset of features from a given dataset [58, 59]. By leveraging optimization algorithms and statistical measures, RapidMiner’s optimized selection functionality aims to enhance both the efficiency and efficacy of predictive models. The process of optimized selection involves iteratively evaluating different feature subsets and assessing their impact on the model’s performance [60]. The operator as shown in Fig. 5, implements two deterministic greedy feature selection algorithms: “forward selection” and “backward elimination.”.

The goal of the forward selection algorithm is to generate the most effective subset of features while disregarding irrelevant and insignificant ones [61–63]. It begins by creating an initial population of n individuals, where n represents the number of attributes in the input Example Set. Each individual in the population uses only one feature. The attribute sets are then evaluated, and the top k sets are selected based on their performance. For each of the k selected sets, the algorithm proceeds as follows: If there are j unused attributes, j copies of the attribute set are made, and exactly one previously unused attribute is added to each copy of the set. The algorithm continues to the next step as long as there has been an improvement in performance in the last p iterations. The Backward Elimination technique begins with an attribute set that includes all features [64, 65]. It evaluates all attribute sets and chooses the top k sets based on their performance. For each of the selected k sets, the algorithm proceeds as follows: If there are j attributes currently used, j copies of the attribute set are made, and exactly one previously used attribute is removed from each copy of the set. The algorithm continues to the next step as long as there has been an improvement in performance in the last p iterations.

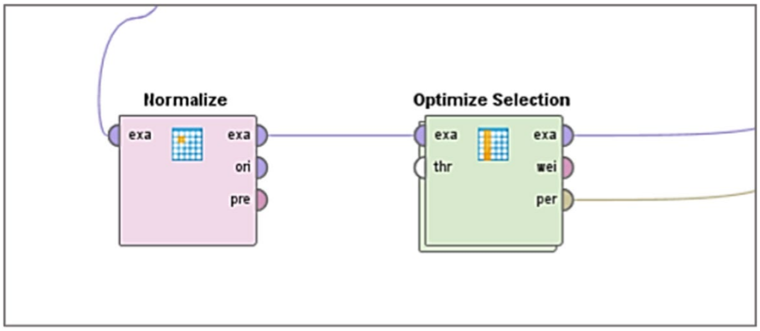
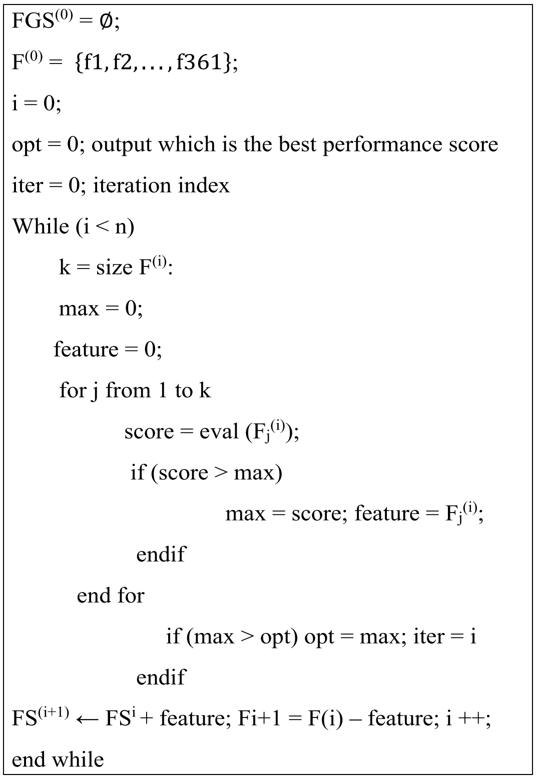


Fig. 5 Normalize and Optimize selection operators in Rapid Miner

Pseudocode of Forward Greedy Search (FGS) Feature Selection:



```
$FGS^{(0)} = \emptyset$;
$F^{(0)} = \{f1, f2, \ldots, f361\}$;
i = 0;
opt = 0; output which is the best performance score
iter = 0; iteration index
While (i < n)
    k = size $F^{(i)}$:
    max = 0;
    feature = 0;
    for j from 1 to k
        score = eval ($F_j^{(i)}$);
        if (score > max)
            max = score; feature = $F_j^{(i)}$;
        endif
    end for
        if (max > opt) opt = max; iter = i
        endif
$FS^{(i+1)} \leftarrow FS^i$ + feature; Fi+1 = F(i) – feature; i ++;
end while
```

Details regarding the parameters of the operators employed in RapidMiner are in Table 2.

Before feature reduction, machine learning models often face challenges such as high dimensionality and redundant or irrelevant features [66–68]. These issues can negatively impact both accuracy and execution time. With a large number of features, models may struggle to extract meaningful patterns from the data, leading to overfitting or poor generalization. Additionally, the computational complexity of training and inference increases significantly with the increasing number of features. However, after feature reduction techniques were applied, such as dimensionality reduction or feature selection, the models experienced improved performance in terms of accuracy as shown in Fig. 6, and execution time as shown in Fig. 7.

Tables 3 and 4 present a summary of the application of various deep learning and machine learning techniques on the TCGA LIHC clinical variables dataset for predicting hepatocellular carcinoma (HCC). This summary includes the performance of these techniques both before and after feature reduction methods were applied. The algorithms utilized in this study encompassed Naive Bayes, Neural Network, Decision Tree, SVM, and KNN. The primary focus of the evaluation was on the prediction of HCC nodules. The results indicate that both the deep learning models and machine learning models exhibited outstanding performance after the implementation of feature reduction methods.

Table 2 Information about the parameters of used operators in RapidMiner

Used operators	Parameters	
Set role	Attribute	Tumor_stauts
	Role	Label
Replace missing values	Replacement value	Average
Weight by information gain	Normalize weight	True
	Sort weights	True
	Sort direction	Ascending
Weight by relief	Number of neighbors	10
	Sample ratio	1.0
Nominal to numerical	Coding type	Dummy coding
Select by weight	Weight relation	Greater equals
	Weight	0.1
Split data	Partitions	Ratio:0.8–0.2
	Sampling type	Stratified sampling
Normalize	Method	Range transformation
	Min	0
	Max	1.0
Optimize selection	Selection direction	Forward
	Max Number of generations	Naive Bayes(6),decision tree (7),Neural nets(6),SVM(4),KNN(7)
Naive Bayes	Laplace correction	True
Decision tree	Criterion	Gain ratio
	Maximal depth	10
	Confidence	0.1
	Minimal gain	0.01
	Minimal leaf size	2
	Minimal size for split	4
	Number of pre-pruning alternatives	3
KNN	K	1
	Measure type	Mixed Euclidean Distance
Neural network	Training cycles	200
	Learning rate	0.01
	Momentum	0.9
SVM	Kernel type	Polynomial
	Kernel degree	2.0
	Kernel cache	200
	Max iteration	100,000
	C	10
	Convergence epsilon	0.001

Before feature reduction, our Neural Network model lumbered through training, achieving an accuracy of 76.00% at the cost of a sluggish 5 min. This sluggishness stemmed from the model struggling to navigate the complexities of a high-dimensional feature space, often getting tangled in irrelevant or redundant information. However, after applying feature reduction techniques, the model shed its excess baggage, emerging lean and mean. It effortlessly soared through training, achieving a remarkable 96% in a mere 1 min and 10 s. This drastic improvement is a testament to the power of feature reduction. By eliminating noisy and superfluous features, we cleared the path for

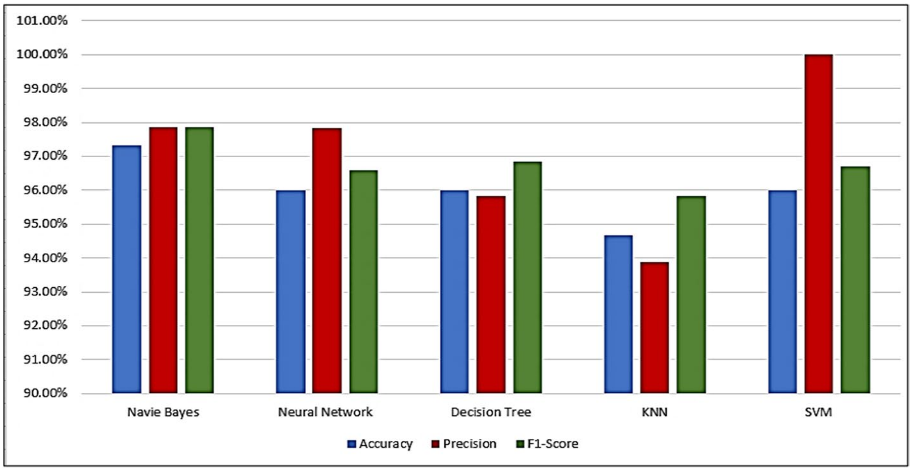


Fig. 6 Performance of used algorithms for HCC Prediction, on the TCGA LIHC clinical variables dataset after feature reduction methods

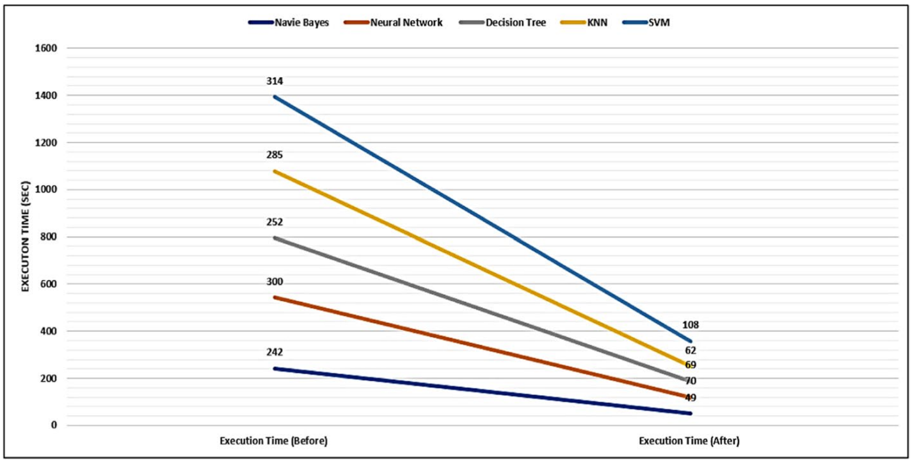


Fig. 7 Execution time of used algorithms for HCC prediction, before and after feature reduction methods in seconds

the model to focus on the truly meaningful relationships within the data, resulting in a more accurate and efficient learning process. This optimization paves the way for faster real-time predictions, reduced computational costs, and ultimately, a more robust and deployable model.

Applying feature reduction techniques to the Naive Bayes model yields notable enhancements in both accuracy and execution time. Specifically, the model achieves an impressive accuracy rate of 97.33%. Moreover, the execution time is significantly reduced to a mere 49 s, showcasing the model's enhanced efficiency in processing and making predictions. These improvements highlight the effectiveness of feature reduction in optimizing the Naive Bayes model's performance, resulting in superior accuracy and faster execution times.

Before implementing feature reduction, the Decision Tree model attains a commendable accuracy of 90.67% but necessitates a relatively lengthy execution duration of 4 min

Table 3 Performance comparison when using each of the deep learning and machine learning algorithms for HCC Prediction, on the TCGA LIHC clinical variables dataset Before Feature Reduction Methods

Approach	Prediction criteria	TRUE (WITH TUMOR)	TRUE (TUMOR-FREE)	Precision (%)	Accuracy (%)	F1-score (%)	Execution time
Naive Bayes	Pred. WITH TUMOR	24	3	88.89%	90.67%	92.63%	4 min and 2 s
	Pred. TUMOR FREE	4	44	91.67%			
	Recall (%)	85.71%	93.62%				
Neural Network	Pred. WITH TUMOR	13	3	81.25%	76.00%	83.02%	5 min
	Pred. TUMOR FREE	15	44	74.58%			
	Recall (%)	46.43%	93.62%				
Decision Tree	Pred. WITH TUMOR	21	0	100%	90.67%	93.05%	4 min and 12 s
	Pred. TUMOR FREE	7	47	87.04%			
	Recall (%)	75%	100%				
SVM	Pred. WITH TUMOR	23	1	95.83%	92%	93.56%	5 min and 14 s
	Pred. TUMOR FREE	5	46	90.20%			
	Recall (%)	82.14%	97.87%				
KNN	Pred. WITH TUMOR	18	0	100%	86.67%	90.38%	4 min and 45 s
	Pred. TUMOR FREE	10	47	82.46%			
	Recall (%)	64.29%	100%				

and 12 s. Nevertheless, following the application of feature reduction techniques, the model undergoes noteworthy enhancements. It accomplishes an impressive accuracy rate of 96%, demonstrating improved precision when classifying instances. Furthermore, the execution time is significantly reduced to a mere 1 min and 9 s. These enhancements underscore the efficacy of feature reduction in optimizing the performance of the Decision Tree model, leading to substantially higher accuracy and faster execution. Moreover, both the SVM and KNN models exhibit superior accuracy, with the SVM model achieving 96.00% accuracy and the KNN model achieving 94.67% accuracy. Notably, the execution times for these models are 1 min and 48 s for SVM and 1 min and 2 s for KNN, respectively.

Discussion

A multitude of machine-learning algorithms have been developed for the prediction of hepatocellular carcinoma. The study [69] explores using a combination of machine learning techniques (ensemble learning) to predict how long Hepatocellular

Table 4 Performance comparison when using each of the deep learning and machine learning algorithms for HCC Prediction, on the TCGA LIHC clinical variables dataset After Feature Reduction Methods

Approach	Prediction criteria	TRUE (WITH TUMOR)	TRUE (TUMOR-FREE)	Precision (%)	Accuracy (%)	F1-score (%)	Execution Time
Naive Bayes	Pred. WITH TUMOR	27	1	96.43%	97.33%	97.87%	49 s
	Pred. TUMOR FREE	1	46	97.87%			
	Recall (%)	96.43%	97.87%				
Neural network	Pred. WITH TUMOR	27	2	93.10%	96%	96.59%	1 min and 10 s
	Pred. TUMOR FREE	1	45	97.83%			
	Recall (%)	96.43%	95.4%				
Decision tree	Pred. WITH TUMOR	26	1	96.30%	96%	96.83%	1 min and 9 s
	Pred. TUMOR FREE	2	46	95.83%			
	Recall (%)	92.86%	97.87%				
SVM	Pred. WITH TUMOR	28	3	90.32%	96.00%	96.70%	1 min and 48 s
	Pred. TUMOR FREE	0	44	100%			
	Recall (%)	100%	93.62%				
KNN	Pred. WITH TUMOR	25	1	96.15%	94.67%	95.83%	1 min and 2 s
	Pred. TUMOR FREE	3	46	93.88%			
	Recall (%)	89.29%	97.87%				

Carcinoma (HCC) patients will survive. The model considers various factors that might influence survival, including patient location, risk factors, and details from clinical trials.

The researchers test fifteen different models, each involving data cleaning, reducing unnecessary features, and then classifying patients based on their predicted survival time. To identify the most important factors, they use four methods: LASSO regression, Ridge regression, a Genetic Algorithm, and a Random Forest. Only the most influential factors are used for prediction.

The models they build include variations of Nu-Support Vector Classification, Ridge Classification (RCV), and Gradient Boosting Ensemble Learning (GBEL), each combined with either L1 or L2 regularization or optimized by a Genetic Algorithm or Random Forest. These models are evaluated based on how accurately they predict survival, using metrics like accuracy, sensitivity, and Area Under the Curve (AUC).

Their findings show that the RFGBEL model (Random Forest combined with Gradient Boosting Ensemble Learning) performs best compared to the others. This model achieves an accuracy of over 93% and a high AUC score of 0.932, indicating strong prediction capabilities. Finally, they compare their RFGBEL model to existing methods and demonstrate its superior ability to predict HCC patient survival.

Also, researchers in the study [70] propose a new NCA-GA-SVM model for predicting HCC survival. This model combines known high-performing techniques (NCA, GA) to improve SVM classification. It achieved high accuracy (96.36%) on a dataset of 165 patients.

This study [71] developed a highly accurate model for diagnosing liver cancer (HCC) that leverages a combination of personalized biological pathways and machine learning. The model achieved exceptional performance in internal testing (AUROC > 0.98) and demonstrated good generalizability to external data. These results suggest this model has great potential for real-world application in HCC diagnosis. Kiani et al. [72] used a microscopic image from the TCGA dataset and utilized a convolutional neural network (CNN) tool named the "Liver Cancer Assistant," it accomplished precise discrimination between hepatocellular carcinoma (HCC) and cholangiocarcinoma. Notably, the model achieved a diagnostic accuracy of 0.885, highlighting its efficacy in accurately identifying and distinguishing between these two distinct forms of liver cancer.

In a study conducted by Wang et al. [73], a deep learning technique involving a convolutional neural network (CNN) was utilized to automate the identification and classification of individual nuclei in tissue images. The CNN was trained using H&E-stained tissue sections of hepatocellular carcinoma (HCC) tumors from the TCGA dataset. Subsequently, a process of feature extraction was carried out, resulting in the identification of 246 quantitative image features. Using an unsupervised learning approach, a clustering analysis was performed, which yielded intriguing results. Surprisingly, this analysis unveiled the existence of three distinct histologic subtypes within the HCC tumors. Importantly, these subtypes were found to be unrelated to previously established genomic clusters and exhibited different prognoses. This study demonstrated the potential of CNN-based image analysis in revealing unique histologic subtypes, offering valuable insights into the prognosis of HCC tumors. Table 5 displays a collection of models proposed by different authors, which have been applied to various HCC-related problems using the TCGA dataset. Table 5 represents the Studies of patients with hepatocellular carcinoma based on the TCGA LIHC dataset.

In this work, we proposed an approach that aims to improve the prediction of hepatocellular carcinoma (HCC) through a comprehensive approach that involves multiple strategies. These strategies include reducing the number of features used in the prediction model through methods such as analyzing feature importance, exploring hidden feature correlations, and employing various algorithms for HCC prediction using clinical variables. We utilized TCGA LIHC clinical variables but the data needed to be cleaned to address any inconsistencies, missing values, or errors. Then the data was formatted and prepared for further analysis which involved scaling the data to a common range, encoding categorical variables, or performing feature engineering to create new features from existing ones. After identifying the optimized feature subset, the dataset was split into two sets: a training set with 301 examples and a testing set with 75 examples. This division was performed using a sampling technique called "Stratified sampling." This sampling technique ensures that random subsets are created while maintaining the consistent distribution of classes within

Table 5 Studies of patients with hepatocellular carcinoma based on the TCGA LIHC dataset

Study	Dataset	Algorithm	Year	Accuracy
Deng et al. [74]	TCGA and HCCDB18 datasets	Unsupervised consistent clustering method	2022	Comparison of Glycolysis and Cholesterol Gene Expression in Normal and Tumor Samples
Cheng et al. [75]	TCGA-LIHC data set	Cox regression analysis	2022	AUC values of the patient's 3-year and 5-year Overall Survival were 0.783 and 0.828, respectively,
Yamashita et al. [76]	Stanford-HCCDET; TCGA	Convolution neural network	2021	The AUROC for tumor tile classification was 0.952 (95% CI 0.948, 0.957) on the internal test set
Saillard et al. [77]	French center and TCGA	Convolution neural network	2020	These CNN-based models demonstrate superior performance compared to traditional models, achieving a C-index ranging from 0.75 to 0.78
Tohme et al. [78]	TCGA-LIHC	ANN	2021	The artificial neural network (ANN) identified a set of 15 genes that exhibited a normalized importance greater than 50%
Kiani et al. [72]	TCGA	CNN	2020	By employing a CNN-based tool, classifying between hepatocellular carcinoma and cholangiocarcinoma exhibited a diagnostic accuracy rate of 0.885
Liao et al. [22]	TCGA and a center in China	Convolution neural network	2020	The predictions of mutations were surpassing an Area Under the Curve (AUC) value of 0.70
Wang et al. [73]	TCGA–LIHC	Convolution neural network	2020	The model demonstrated high accuracy, achieving an overall classification rate of 99% for tumor cells and 97% for lymphocytes
Shi et al. [79]	1 center in China; TCGA	Convolution neural network	2021	The deep learning-based "stratifies the study population into five groups with distinct prognoses in both the Zhongshan cohort ($p < 0.0001$) and TCGA cohort ($p = 0.0003$)"

those subsets, aligning with the overall class distribution in the entire dataset. In other words, Stratified sampling helps to preserve the proportional representation of different classes during the creation of training and testing sets, which is essential for maintaining the integrity of the dataset and ensuring reliable model evaluation. The application of feature reduction techniques to the Naive Bayes model leads to significant improvements in accuracy and execution time. With these techniques implemented, the model achieves an impressive accuracy rate of 97.33%. Additionally, the execution time is drastically reduced to just 49 s, demonstrating the enhanced efficiency of the model in processing and making predictions. These enhancements clearly illustrate the effectiveness of feature reduction in optimizing the performance of the Naive Bayes model, resulting in higher accuracy and faster execution times.

Limitations

Although machine learning and deep learning have shown promise in various medical applications, including hepatocellular carcinoma (HCC) prediction, there are several limitations associated with their use in this context.

One major limitation is the requirement for large and high-quality datasets. Machine learning algorithms, including deep learning models, heavily rely on vast amounts of well-curated data to learn patterns and make accurate predictions. However, acquiring such datasets for HCC prediction can be challenging due to the rarity of the disease and the need for comprehensive clinical and imaging data. The limited availability of annotated HCC datasets hampers the development and evaluation of robust models.

Interpretability and explainability are crucial in medical decision-making, and this is another limitation of the deep learning model. While these models have demonstrated remarkable predictive capabilities, they often function as black boxes, making it difficult to understand the underlying reasons behind their predictions. This lack of interpretability raises concerns in medical settings, where clinicians need to have confidence in the decision-making process and understand the factors contributing to a prediction.

The generalizability of machine learning and deep learning models can also be a limitation. Models trained on specific populations or datasets may not perform as well when applied to different patient populations or settings. The heterogeneity of HCC, including variations in tumor characteristics, genetic profiles, and patient demographics, can introduce challenges in developing models that can effectively predict HCC across diverse populations. Furthermore, the potential for bias in machine learning models is another limitation. Biases can be introduced during the data collection process, such as underrepresentation of certain demographic groups or confounding factors. If the models are trained on biased datasets, they may perpetuate or even amplify existing biases, leading to inaccurate predictions and disparities in healthcare outcomes.

Conclusion and future work

In conclusion, this study focused on the prediction of hepatocellular carcinoma (HCC), a prevalent form of liver cancer, using machine learning algorithms. The objective was to assess the effectiveness of feature reduction techniques in enhancing the performance of HCC prediction models. By comparing the performance of various machine learning algorithms on both the original high-dimensional dataset and a reduced feature subset, this study demonstrated that feature reduction significantly improves the accuracy and execution time of HCC prediction models. The employed feature reduction techniques, including weighting features, hidden features correlation, feature selection, and optimized selection, helped extract a reduced feature set that captured the most relevant information related to HCC. The experimental results obtained from a comprehensive dataset of clinical features of HCC patients showed that the reduced feature set consistently outperformed the original high-dimensional dataset in terms of prediction accuracy. The decision trees, Naive Bayes, K-nearest neighbors, neural networks, and support vector machines (SVM) algorithms achieved accuracies of 96%, 97.33%, 94.67%, 96%, and 96.00%, respectively, after applying feature reduction techniques. These findings suggest that feature reduction methods can be effectively employed in HCC prediction models, leading to improved accuracy and faster execution times. The application of machine learning algorithms, combined with feature reduction techniques, holds great potential for the early diagnosis and effective treatment of HCC, ultimately improving patient outcomes.

While current models using clinical variables for HCC prediction show promise, there are several areas for future work to improve accuracy, personalize risk assessment, and ultimately guide better patient outcomes. Integrating Multimodal Data by Exploring combining clinical data with other modalities like genetic information, imaging data (MRI, CT scans), and blood-based biomarkers. Deep learning models can be particularly adept at handling such diverse data sources. Also, train and validate models on large, geographically diverse datasets to ensure generalizability and avoid overfitting to specific populations. Account for the presence of other chronic conditions like diabetes or hepatitis that may influence HCC development. Develop models that can incorporate longitudinal data (changes in clinical variables over time) to predict risk changes and identify high-risk patients earlier. By focusing on these future work directions, we can improve the accuracy and clinical utility of HCC prediction models using clinical variables, leading to earlier detection, better risk stratification, and ultimately improved patient outcomes.

Author contributions

This work was carried out in collaboration among all authors. All Authors designed the study, performed the statistical analysis, and wrote the protocol. All Authors managed the analyses of the study, managed the literature searches, and wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The TCGA LIHC clinical data set is publicly available.
Data Access Points: There are a couple of resources where researchers can access the data: GDC Data Portal: "https://portal.gdc.cancer.gov/". TCIA: "https://imaging.cancer.gov/informatics/cancer_imaging_archive.htm".

Declarations

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication

All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 27 February 2024 Accepted: 31 May 2024
Published online: 18 June 2024

References

1. Torre LA, et al. Global cancer statistics, 2012. CA Cancer J Clin. 2015;65(2):87–108.
2. DeWaal D, et al. Hexokinase-2 depletion inhibits glycolysis and induces oxidative phosphorylation in hepatocellular carcinoma and sensitizes to metformin. Nat Commun. 2018;9(1):446.

3. Santos MS, et al. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform. 2015;58:49–59.
4. Ali L, Bukhari S. An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. Irbm. 2021;42(5):345–52.
5. Książek W, et al. A novel machine learning approach for early detection of hepatocellular carcinoma patients. Cogn Syst Res. 2019;54:116–27.
6. Ali L et al. A multi-model framework for evaluating type of speech samples having complementary information about Parkinson's disease. In: 2019 International conference on electrical, communication, and computer engineering (ICECCE). IEEE; 2019.
7. Abdar M, et al. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recogn Lett. 2020;132:123–31.
8. Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Syst Appl. 2014;41(4):1476–82.
9. Shi J, et al. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform. 2017;22(1):173–83.
10. Zhi X, et al. Efficient discriminative clustering via QR decomposition-based linear discriminant analysis. Knowl-Based Syst. 2018;153:117–32.
11. Ali L et al. Early detection of heart failure by reducing the time complexity of the machine learning based predictive model. In: 2019 international conference on electrical, communication, and computer engineering (ICECCE). IEEE; 2019.
12. Ravikulan A, Rostami K. Leveraging machine learning for early recurrence prediction in hepatocellular carcinoma: a step towards precision medicine. World J Gastroenterol. 2024;30(5):424.
13. Hong H, et al. Prediction of hepatocellular carcinoma development in Korean patients after hepatitis C cure with direct-acting antivirals. Gut and Liver. 2024;18(1):147.
14. Abajian A, et al. Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning—an artificial intelligence concept. J Vasc Intervent Radiol. 2018;29(6):850–7.
15. Ioannou GN, et al. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA Netw Open. 2020;3(9):e2015626–e2015626.
16. Nam JY, et al. Deep learning model for prediction of hepatocellular carcinoma in patients with HBV-related cirrhosis on antiviral therapy. JHEP Rep. 2020;2(6): 100175.
17. Nam JY, et al. Novel model to predict HCC recurrence after liver transplantation obtained using deep learning: a multicenter study. Cancers. 2020;12(10):2791.
18. Ali MA, et al. A novel method for survival prediction of hepatocellular carcinoma using feature-selection techniques. Appl Sci. 2022;12(13):6427.
19. Cao Y, et al. Prediction model for recurrence of hepatocellular carcinoma after resection by using neighbor2vec based algorithms. Wiley Interdiscip R Data Min Knowl Discov. 2021;11(2): e1390.
20. Zhang Y, et al. Deep learning with 3D convolutional neural network for noninvasive prediction of microvascular invasion in hepatocellular carcinoma. J Magn Reson Imaging. 2021;54(1):134–43.
21. Zhang Y-B, et al. Development of a machine learning-based model for predicting risk of early postoperative recurrence of hepatocellular carcinoma. World J Gastroenterol. 2023;29(43):5804.
22. Liao H, et al. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. Clin Transl Med. 2020;10(2): e102.
23. Deng Z, et al. Mining TCGA database for tumor microenvironment-related genes of prognostic value in hepatocellular carcinoma. BioMed Res Int. 2019;2019:2408348.
24. Wang K, et al. A novel immune-related genes prognosis biomarker for hepatocellular carcinoma. Aging (Albany NY). 2021;13(1):675.
25. Bannister CA, et al. A genetic programming approach to development of clinical prediction models: a case study in symptomatic cardiovascular disease. PLoS ONE. 2018;13(9): e0202685.
26. Dong Y, et al. A novel surgical predictive model for Chinese Crohn's disease patients. Medicine. 2019;98(46): e17510.
27. Karhade AV, et al. Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation. Spine J. 2019;19(11):1764–71.
28. Scheer JK, et al. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J Neurosurg Spine. 2017;26(6):736–43.
29. Adams S, Beling PA, Cogill R. Feature selection for hidden Markov models and hidden semi-Markov models. IEEE Access. 2016;4:1642–57.
30. Bjaoui M et al. Depth insight for data scientist with RapidMiner «an innovative tool for AI and big data towards medical applications». In: Proceedings of the 2nd international conference on digital tools & uses congress; 2020.
31. Roy SP, Kasat A. Diabetic prediction with ensemble model and feature selection using information gain method. In: 2024 2nd international conference on intelligent data communication technologies and internet of things (IDCIoT). IEEE; 2024.
32. Ihianle IK, et al. Minimising redundancy, maximising relevance: HRV feature selection for stress classification. Expert Syst Appl. 2024;239: 122490.
33. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn. 2003;53:23–69.
34. Shukla AK, et al. Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques. J Intell Fuzzy Syst. 2020;38(5):6637–48.
35. Haq AU, et al. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst. 2018;2018:1–21.
36. Theng D, Bhoyar KK. Feature selection techniques for machine learning: a survey of more than two decades of research. Knowl Inf Syst. 2024;66(3):1575–637.

37. Gao J, et al. Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection. Knowl-Based Syst. 2024;286: 111380.
38. Wang X, Yan Y, Ma X. Feature selection method based on differential correlation information entropy. Neural Process Lett. 2020;52:1339–58.
39. Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl Soft Comput. 2020;97: 105524.
40. Raju VG et al. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 third international conference on smart systems and inventive technology (ICSSIT). IEEE; 2020.
41. Zhou S, et al. Breast cancer prediction based on multiple machine learning algorithms. Technol Cancer Res Treat. 2024;23:15330338241234792.
42. Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recogn Lett. 2001;22(5):563–82.
43. Ajbar W, et al. Development of artificial neural networks for the prediction of the pressure field along a horizontal pipe conveying high-viscosity two-phase flow. Flow Meas Instrum. 2024;96: 102541.
44. Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification, Taipei, Taiwan; 2003.
45. Parashar G, Chaudhary A, Pandey D. Machine learning for prediction of cardiovascular disease and respiratory disease: a review. SN Comput Sci. 2024;5(1):196.
46. Jayalakshmi T, Santhakumaran A. Statistical normalization and back propagation for classification. Int J Comput Theory Eng. 2011;3(1):1793–8201.
47. Acharya UR, et al. Automated diagnosis of glaucoma using texture and higher order spectra features. IEEE Trans Inf Technol Biomed. 2011;15(3):449–55.
48. Snelick R, et al. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. IEEE Trans Pattern Anal Mach Intell. 2005;27(3):450–5.
49. Wen X, et al. Efficient feature selection and classification for vehicle detection. IEEE Trans Circuits Syst Video Technol. 2014;25(3):508–17.
50. Esfahani ET, Wang S, Sundararajan V. Multisensor wireless system for eccentricity and bearing fault detection in induction motors. IEEE/ASME Trans Mechatron. 2013;19(3):818–26.
51. Pan J, Zhuang Y, Fong S. The impact of data normalization on stock market prediction: using SVM and technical indicators. In: Soft computing in data science: second international conference, SCDS 2016, Kuala Lumpur, Malaysia, September 21–22, 2016, Proceedings 2. Springer; 2016.
52. Kadir A et al. Leaf classification using shape, color, and texture features; 2013. arXiv preprint arXiv:1401.4447.
53. Wang C-M, Huang Y-F. Evolutionary-based feature selection approaches with new criteria for data mining: a case study of credit approval data. Expert Syst Appl. 2009;36(3):5900–8.
54. Wu W, et al. Evaluation of normalization methods for cDNA microarray data by k-NN classification. BMC Bioinform. 2005;6:1–21.
55. Liu Z. A method of SVM with normalization in intrusion detection. Procedia Environ Sci. 2011;11:256–62.
56. Su D et al. Anomadroid: profiling android applications' behaviors for identifying unknown malapps. In: 2016 IEEE Trustcom/BigDataSE/ISPA. IEEE; 2016.
57. Peterson RA. Finding optimal normalizing transformations via best normalize. R Journal. 2021;13(1):310–29.
58. El-Hasnony IM, et al. Improved feature selection model for big data analytics. IEEE Access. 2020;8:66989–7004.
59. Song X-F, et al. A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data. IEEE Trans Cybern. 2021;52(9):9573–86.
60. Mohamad M, et al. Enhancing big data feature selection using a hybrid correlation-based feature selection. Electronics. 2021;10(23):2984.
61. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: a review. J King Saud Univ Comput Inf Sci. 2022;34(4):1060–73.
62. Camattari F et al. Greedy feature selection: Classifier-dependent feature selection via greedy methods. arXiv preprint arXiv:2403.05138; 2024.
63. Chen W, Sun X. Dynamic multi-label feature selection algorithm based on label importance and label correlation. Int J Mach Learn Cybern. 2024. https://doi.org/10.1007/s13042-024-02098-3.
64. Habib M, Okayli M. Evaluating the sensitivity of machine learning models to data preprocessing technique in concrete compressive strength estimation. Arab J Sci Eng. 2024. https://doi.org/10.1007/s13369-024-08776-2.
65. Peng M, et al. scFSNN: a feature selection method based on neural network for single-cell RNA-seq data. BMC Genomics. 2024;25(1):264.
66. Ayesha S, Hanif MK, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf Fus. 2020;59:44–58.
67. Ray P, Reddy SS, Banerjee T. Various dimension reduction techniques for high dimensional data analysis: a review. Artif Intell Rev. 2021;54:3473–515.
68. Zebari R, et al. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1(2):56–70.
69. Sharma M, Kumar N. Improved hepatocellular carcinoma fatality prognosis using ensemble learning approach. J Ambient Intell Humaniz Comput. 2022;13(12):5763–77.
70. Książek W, Turza F, Pławiak P. NCA-GA-SVM: a new two-level feature selection method based on neighborhood component analysis and genetic algorithm in hepatocellular carcinoma fatality prognosis. Int J Numer Methods Biomed Eng. 2022;38(6): e3599.
71. Cheng B, Zhou P, Chen Y. Machine-learning algorithms based on personalized pathways for a novel predictive model for the diagnosis of hepatocellular carcinoma. BMC Bioinform. 2022;23(1):248.
72. Kiani A, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Dig Med. 2020;3(1):23.
73. Wang H, et al. Single-cell spatial analysis of tumor and immune microenvironment on whole-slide image reveals hepatocellular carcinoma subtypes. Cancers. 2020;12(12):3562.

74. Deng W, et al. Classification and prognostic characteristics of hepatocellular carcinoma based on glycolysis cholesterol synthesis axis. J Oncol. 2022. https://doi.org/10.1155/2022/2014625.
75. Cheng D, et al. Identification and construction of a 13-gene risk model for prognosis prediction in hepatocellular carcinoma patients. J Clin Lab Anal. 2022;36(5): e24377.
76. Yamashita R, et al. Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Sci Rep. 2021;11(1):1–14.
77. Saillard C, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology. 2020;72(6):2000–13.
78. Tohme S, et al. The use of machine learning to create a risk score to predict survival in patients with hepatocellular carcinoma: a TCGA cohort analysis. Can J Gastroenterol Hepatol. 2021. https://doi.org/10.1155/2021/5212953.
79. Shi J-Y, et al. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. Gut. 2021;70(5):951–61.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.