Research Article

Weighted Synthetic Minority Over-Sampling Technique (WSMOTE) Algorithm and Ensemble Classifier for Hepatocellular Carcinoma (HCC) In Liver Disease System

K. Jeyalakshmi[a] and R. Rangaraj[b]

[a]Associate Professor ,Department of Computer science , Hindustan College of Arts and Science, Coimbatore, Tamilnadu. E-mail: Jeyas1201@gmail.com
[b]Professor and Head, PG and research, Department of Computer science, Hindustan College of Arts and Science Coimbatore, Tamilnadu.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: HCC (Hepato Cellular Carcinoma) is a generic liver cancer causing death in people suffering from LC (Liver Cirrhosis). Early prognosis of HCC is a significant factor to the line of LC treatment for clinicians. Though current treatments for HCC have been effective, patients have responded negatively or exhibited aggressive biological behaviours. Identification accuracy reduces when sub-optimal models of MLTs (Machine Learning Techniques) process multiple classes. MLT models base their predictions on their training phases where class imbalances in datasets have retained challenges in terms of unsatisfactory results. This research work attempts to overcome these issues with its proposed WSMOTE (Weighted Synthetic Minority Over-sampling Technique) algorithm which is targeted at dataset imbalances. Missing values are imputed using IFCM (Improved Fuzzy C Means) clustering for enhancing analysis accuracy. Imputed features are chosen selectively using IEFS (Intelligence Ensemble-Based Feature Selection). A heterogeneous ensemble classifier using Bootstrap aggregation is applied for combining predictions of multiple classifiers including KSVMs (Kernel Support Vector Machines) and FCNNs (Fuzzy Convolution Neural Networks) for accuracy in outputs. This work's schema when tested on MATLAB (Matrix Laboratory) was found to classify better than most other methods in terms of precision, recall, F-measure and accuracy.

Index Terms: Liver disease prediction, Hepatocellular Carcinoma (HCC), missing data imputation, clustering, Weighted Synthetic Minority Over-sampling Technique (WSMOTE), Dimensionality reduction, Feature Selection (FS), Intelligence Ensemble-Based Feature Selection (IEFS), Ensemble classifier, and Fuzzy Convolutional Neural Network (FCNN).

1. Introduction

Cancers have become a leading cause of global deaths where LCs belong to life-threatening disease category and have increased phenomenally (Indhumathy et al., 2018); (Villanueva, 2019). HCC is the 5th most common type of cancer, accounting to ninety percent of LCs and ranks 2nd in cancer related mortality. Primary LCs includes HCC (75–85%), intrahepatic cholangiocarcinoma (10–15%) while metastatic cancers of other human parts account to 10%. Several factors contribute towards Cirrhosis in HCC: HBV (Hepatitis B Virus)/ HCV (Hepatitis C Virus) infections; intense alcohol consumptions; obesity; type 2 diabetes and contaminated aflatoxin food intakes (Bray et al., 2018).

HCC has recorded annual incidence of 2–7% in HCV/HBV patients with liver cirrhosis in America. With increases in highly effective and well directed antiviral agents for HCV treatments, cured patients percentage has also increased. Global increases in obesity have also resulted in increased NAFLDs (Non-alcoholic Fatty Liver Diseases), the current fastest CLDs (Chronic Liver Diseases), HCC and cirrhosis (Kanwal & Singal, 2019).

Diagnostics based on risk scores obtained from signature gene expressions (Zhou et al., 2017);(Qu et al., 2019) are highly sensitive to measurements (Guan et al., 2016), but are hardly used in clinical prognosis. Thus, it becomes imperative to design new DMTs (Data Mining Techniques) for early identification of HCC as cancerous tissues can engulf other non- cancerous tissues and transform their molecular characteristics into cancerous parts. Healthcare automated systems or CDSSs (Clinical Decision Support Systems) can be of great assistance to clinicians in timely and accurate diagnosis of diseases by using patients past clinical history. CDSSs use MLTs to enhance medical decision quality while minimizing diagnostic errors. MLTs learn from experience gained from past clinical history of patients and thus contribute towards diagnostic error reductions (Shimizu et al., 2018). The use of DMTs have increased substantially and are being applied widely in healthcare diagnosis and

assessments including cardiac disorders, LCs, BCs (Breast Cancers), Parkinson disease and Alzheimer's disease to name a few. Thus, it becomes imperative to design new DMTs for early identification of HCC. Many studies have proposed schemes for identifying LCs and HCC. MLTs predictions of HCC from datasets have to overcome several issues in recorded data like missing values, data imbalances (López et al., 2013); (Ali-Gombe, A & Elyan, 2019); (Stefanowski, 2016) and high dimensionality of features.

Imbalances are an issue for both binary and multi-class data. Imbalances in data imply that one of classes is more or the number of observations is not equal amongst classes impacting classifications. Oversampling is used when the quantity of data is insufficient and helps increase the size of samples. Data imbalance effects can be nullified using oversampling for creating balanced datasets. SMOTE (Synthetic Minority Over-sampling Technique) is the main technique used in data oversampling as it normalizes dataset imbalances by reducing the size of the class which is more in numbers.

Datasets generally have large number of features belonging to multiple classes which may be irrelevant to MLT classifications. Since, these irrelevant features hamper classification model performances, feature selections play a significant role for enhancing classifier's accuracy. This study is aimed at finding the best subset of features and evaluate predictive classification performance. WSMOTE algorithm introduced in this work solves dataset imbalance issue with IEFS Feature selection for the LC classifications. The efficiency of predictions in enhanced further by using multitude of classifiers for classifying samples as positive or negative to LCs. The combination of KSVM and FCNNs evaluate the generated feature subsets on LC datasets obtained from UCI (University of California, Irvine) repository.

2. Literature Review

Chen et al., (2020) proposed data that a real clinically diagnosed HCC patient was collected from a University of Coimbra and Coimbra Hospital in Portugal, and separated the data into testing and training to predict the death of HCC and find out the key factors from the prediction model. The prediction model includes Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR). The results of this work showed that the G-means of the three modelling methods are 0.76 (LR), 0.72 (DT), and 0.68 (SVM). The best performance is Logistic Regression (LR), and find out the key factors that affect the survival rate of HCC include Aspartate transaminase (U / L), Age at diagnosis, and Alkaline phosphatase (U / L).

The study by Dong et al., (2019) used Cox regression clubbed with SVM-RFE (Support Vector Machine-Recursive Feature Elimination) and FW-SVM (Forward-SVM) algorithms where 47,099 differentially methylated sites were screened. The study's model assessed risk into three categories namely high, intermediate and low based on 134 methylated sites for overall survival of patients. A 10-fold cross-validation of the proposed model had a score of 0.95 with satisfactory predictions where 26/33 samples were classified accurately while using the testing set.

SVMs were used by Sato et al., (2019) in their study for creating a predictive model of single tumor marker. The proposed scheme reduced misclassifications by almost 50%. The study's proposed framework used grid search for automatically selecting the best predictive classifier and its corresponding hyper-parameter. Important variables which can discriminate predictive classes was assessed using Gini impurity decrements. The study's framework when applied on various kinds of data showed promising results and that it was a potential candidate for furthering in academic researches and clinical practices.

MLTs were also proposed in the study by Gui et al., (2015). Their approach was based on mRMR (maximum-Relevance–Minimum-Redundancy) algorithm followed by IFS (Incremental Feature Selection). The study used microarray data generated from 95 samples which included 43 tumorous and 52 non-tumorous samples. The study identified 117 gene probes for optimal separation of samples. A molecular interaction network based on PPIs (Protein–Protein Interactions) constructed from STRING database data identified 187 genes. The study's analysis revealed the role of ubiquitin C in HCC pathogenesis. Their analysis based on GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) showed that their sub-network's identifications could enriched in biological processes related to cell deaths and thus bringing new insights into HCC's understanding.

Recurrence of HCC was predicted using DMTs by Iwahashi et al., (2020) in their study which focused on HCC operated patients. The study used ADTs (Alternating Decision Trees) validated using a 10-fold cross validation process. 323 HCC patients with hepatic resectioning were included in their study's Clinicopathological data which had 156 cancer recurrence patients. The study examined the usefulness of DMTs on the predictability of HCC postoperative recurrences in analysis with satisfactory results.

Neighbor2vec based algorithm was used by Cao et al., (2021) in their novel prediction model designed for HCC recurrence. Their model worked in three phases. Their preparation stage used PCCs (Pearson Correlation Coefficients) for exploring HCC recurrence's independent predictors. For low correlations between individual dimensions, KNNs (K-Nearest Neighbors) found prediction target for each patient (neighbor2vec). In the third phase, the obtained vectors lists were input into MLTs like LRs (Logistic Regressions), KNNs, DTs (Decision Trees), DNNs (Deep Neural Networks) and NB (Naive Bayes) for establishing the study's neighbor2vec prediction model. Their experimentations on real data from China's Shandong Provincial Hospital using proposed neighbor2vec based prediction model outperformed other benchmarked models, specifically NB. The study's proposed model achieved an accuracy of 83.02%, recall 82.86% and 77.6% in precision.

ANNs (Artificial Neural Networks) were used in evaluating predictors for HCC mortality by Chiu et al., (2013) in their work. The study compared LR and ANN models with significant predictors on patient's survival and undergoing HCC resections based on their predictive accuracies. Their prognostic model constructed on 434 patient data used 21 input features obtained by Cox regression model. The study evaluated both significant predictors and their accuracy in predictions. Their experimentations proved ANNs utility in identifying predictors for mortality with accurate predictions when compared to conventional methods. The study suggested use of DMTs as supplementary tools for clinical decisions in prognostic evaluations of physicians.

mRMR was also used for selecting features by Zhang et al., (2020) in their proposal. The study obtained 11-gene-pair by its incremental feature selections. The obtained gene signatures were also applied on independent datasets for evaluating identification of HCC. The proposed scheme's computations discriminated HCC from neighbouring non-cancerous tissues accurately even in inaccurate specimen samples and for minimum biopsy specimens. Their methodology proved its utility in practical and effective HCC diagnosis.

PLA (Pyogenic Liver Abscess) from clinical data was used to assess cancers by Hon et al., (2020) in their study. The study aimed to assist clinicians in early assessment of cancers. The study used Binary LRs for determining Ors (Odds Ratios) and CIs (Confidence Intervals) of 95%. The model constructed cancer classifications using optimized risk factors for cancer with C5.0, a DT which was also compared with different models in terms of accuracy. The results displayed the model's supremacy in cancer predictions.

A novel MLT was proposed by Książek et al., (2019) in their study to identify HCC with 165 patient's data. The study normalized data as a pre-processing step. GA (Genetic Algorithm) with stratified 5-fold cross-validation was used two times. Initially GA optimized parameters and then again selected features. This work also used SVMs (type C-SVC) with a dual level genetic optimizer in training. The study's feature selections resulted in high accuracy score of 0.8849 with 0.8762 as its F1-Score.

Data imputations were the primary proposal in the study of Santos et al., (2015). The study imputed data using HEOM which used corresponding distance metrics for Heterogeneous as well as missing data while K-means identified patient groups by clustering them. The study managed to reduce the impact of patient profiles in processing by minimizing survival prediction data. The proposal used SMOTE and K-means clustering to generate a representative dataset and trained with multitude of MLTs including LRs and NNs (Neural Networks). The study's schema of NN usage outperformed other techniques, suggesting classical approach enhancements used for HCC predictions.

APOs (Artificial Plant Optimizations) were used for HCC predictions in the study of Divya & Radha, (2019). Their efficient sampling approach overcame class imbalances used IRUS (Inverse Random under Sampling). IRUS created distinct partitions where minority and majority class samples were demarcated by boundaries. Their algorithmic optimization selected optimal features for

improving classifier efficiency and effectiveness. The selected optimal features were used for classifying patients with or without HCC. SVMs and RFs (Random Forests) were used in classifications. The study's schema proved its effectiveness in experimentations in terms of accuracy, specificity, sensitivity and balanced accuracy when benchmarked with other techniques.

3. Proposed Methodology

This research work uses DMT's data imputations, dimensionality reductions, feature selections, balancing imbalanced datasets with oversampling, classification and result evaluations. This proposed work uses WSMOTE algorithm for managing imbalanced datasets while ensemble classifiers FCNNs and KSVMs classify data. The flow of this research work is depicted as figure 1.

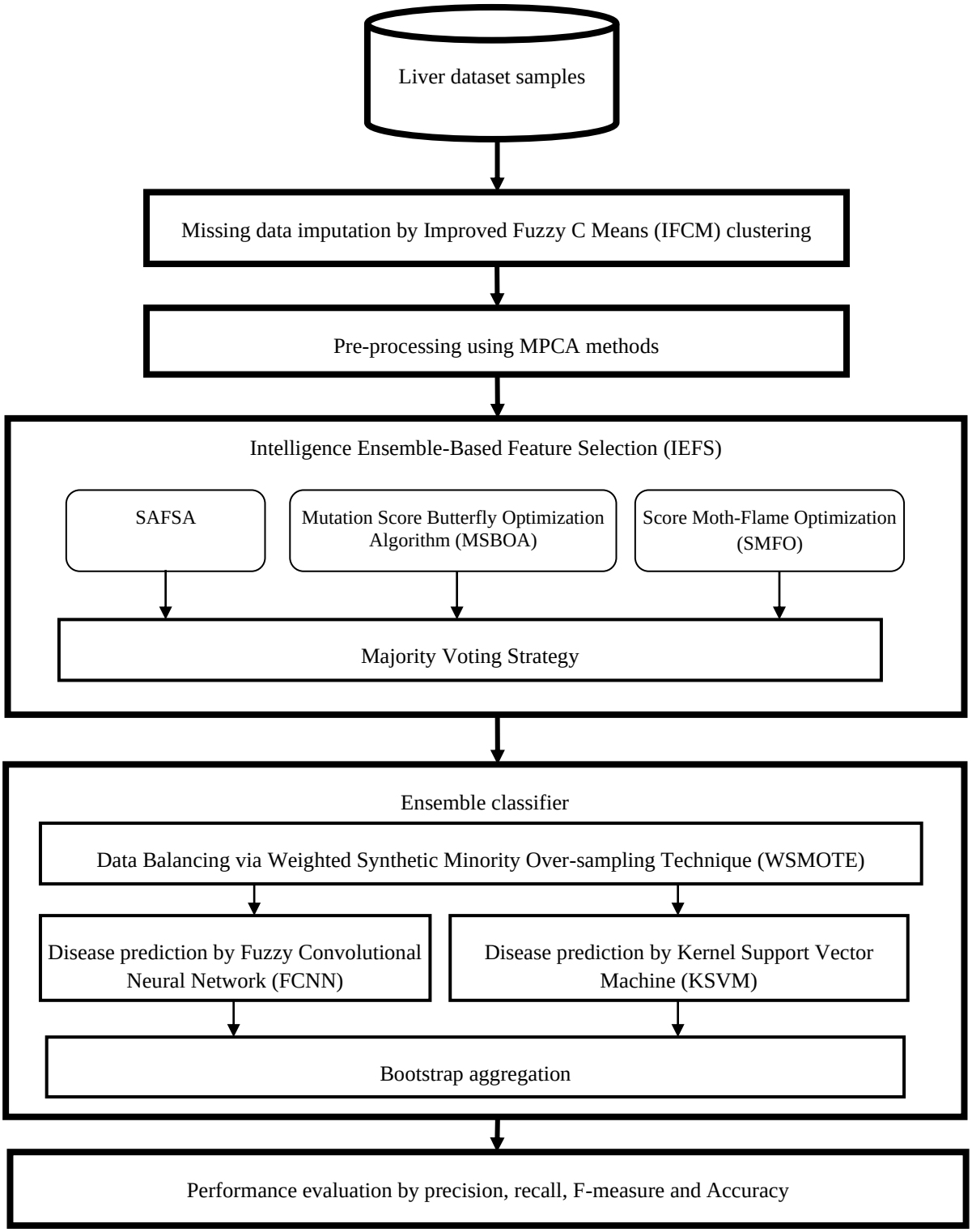

Figure 1. Overall Flow of the Proposed System

3.1. Dataset Description

Three benchmark datasets from the UCI's (University of California, Irvine) ML (Machine Learning) repository webpage have been used for implementations.

Indian Liver Patient Records (ILPD): The use of this dataset was aimed at helping doctors. The dataset has 167 non-liver and 416 liver disease records collected from Andhra Pradesh patients in India. It is collected from https://archive.ics.uci.edu/ml/datasets/ILPD+ (Indian+Liver+Patient+Dataset). The dataset's class labels divide records into disease or no disease samples. Further, there are 142 female and 441 male patient records and patients exceeding the age of 89 are marked 90 (Refer Table 1).

Hepatocellular Carcinoma (HCC) dataset is from University of California, Irvine (UCI) Machine Learning (ML) repository webpage from https://archive.ics.uci.edu/ml/datasets/HCC. The dataset includes both missing values and imbalance nature of class label. Here the missing values were imputed by Improved Fuzzy C Means (IFCM).

HCC Balanced Dataset: In this study, the HCC balanced dataset samples have 50 attributes with 204 instances which is collected from https://github.com/amazzocchi13/HCC-Prediction-Model-ML. Out of 204 cases, 102 cases labeled as "lives (No), 102 as "dies (Yes). Table 1 shows the dataset characteristics.

HCC Survival Dataset: This dataset has 165 HCC patients and includes both missing values and imbalance nature of class labels. Here the missing values were imputed using IFCM. HCC patient's data, formed from a Portugal University Hospital encompasses risks, demographics, laboratory and survival features from HCC dataset from https://archive.ics.uci.edu/ml/datasets/HCC+Survival. Forty nine survival features suggested by EASL-EORTC (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer) Clinical Practice Guidelines are used in this work with twenty three quantitative and twenty six qualitative features. Further, the dataset has 10.22% missing values and only eight patient's complete information is provided. The target variable, 1 year survival is encoded as binary (0- Dead, 1 – Alive). Table 1 details on the dataset's characteristics.

Attributes Description of ILPD: ILPD dataset consists a total of 10 attributes which includes numerical values, one as a class label ["lives (No), "dies (Yes)]. They are categorized in Nominal, and Category.

Attributes Description of HCC: The HCC dataset consists a total of 50 attributes which includes 26 qualitative variables + 23 quantitative variables (referred as predictable attribute or input attributes), one as a class label ["lives (No), "dies (Yes)]. They are categorized in Nominal, Continuous, Ordinal and Integer.

Gene Expression Dataset: The gene expression dataset is collected from CancerLivER. Three gene expressions such as GSE102079, GSE107170, and GSE25097 have been used for implementation. GSE102079 includes of 257 patients with 22048 gene expression microarray (Robust Multi Array (RMA)) for Hepatocellular Carcinoma (HCC), GSE107170 includes of 307 patients with 22048 gene expression microarray(RMA) for HCC, GSE25097 includes of 557 patients with 22048 gene expression microarray(RMA) for HCC(See table 1).

Table 1. Dataset Characteristics

Datasets	**Attributes**	**Instances**	**Missing Values**
University of California, Irvine (UCI)- HCC Survival Dataset	50	165	Yes
HCC balanced dataset	50	204	No
Indian Liver Patient Records	11	583	No
Gene expression dataset	22048	1121	No

3.2. Data Preprocessing

Improved Fuzzy C Means (IFCM) clustering is an efficient way to estimate missing values of datasets. Missing values are assigned to candidate values and computed using fuzzy membership function. Mathematically assuming, X is the raw data matrix including missing values where $X = \{x_1, \ldots x_k\}$ and n is the count of samples. IFCM clustering is performed based on the squared Euclidean distance between any 2 data points $\left|\left|x_k - y_{ji}\right|\right|^2$ (Huang et al., 2020). If the difference between new membership degree $u(x_k, y_{ji})^*$ and old membership degree $u(x_k, y_{ji})$ is > threshold ε, then new cluster centroid is updated and clustered using IFCM algorithm. In the general FCM clustering algorithm, instead of considering the distance $d(x_k, y_{ji})$ which is replaced with kernel function.

3.3. Modified Principal Component Analysis (MPCA) For Dimensionality Reduction

The liver disease dataset might consist of most noisy and the more irrelevant features. This might increase the computation overhead of the classifier which can be avoided by pre-processing the input dataset. This research work reduces feature dimensionality with MPCA (Modified Principal Component Analysis) where PCA has the ability to reduce dimensionality in multivariate data and effectively choose important features needed for classification of data. The proposed MPCA can reduce errors while de-correlating features. The guarantee that relevant classes (liver disease) can be detected by PCA is less and hence this work uses MPCA which constructs three matrices using covariance values, SVDs (Singular-Value Decompositions) and recursions. The differences between the covariance matrix and matrix of whole dataset thus give the dimensionally reduced features.

3.4. Intelligence Ensemble-Based Feature Selection (IEFS)

Wrapper approach that employs Intelligence algorithms, namely, Score based Artificial Fish Swarm Algorithm (SAFSA), Mutation Score Butterfly Optimization Algorithm (MSBOA) and Score Moth-Flame Optimization (SMFO) with accuracy of Fuzzy Convolutional Neural Network (FCNN) classifier as fitness function for feature selections where algorithms select three feature subsets. Thus, in the study optimal features in these feature subsets are fed to FCNN which trains on this data.

3.4.1.Score based Artificial Fish Swarm Algorithm (SAFSA)

Once missing data imputed and dimensionality reduced, it is important to select the most relevant features for increasing the classification accuracy. This optimal feature selection can be done by the SAFSA algorithm which can select the most optimal features from the given input dataset. Information gain and classification accuracy are fitness function in this study. SAFSA is inspired by collective movements and social behaviour of fishes. Their social behaviours include food searches, migrations and handling unforeseen dangers. Their interactions are a result of their intelligent social behaviour. SAFSA has better classification accuracy based on its feature selections. The introduction of random behaviour in AFSA for finding global optimal solutions, finishing of the iteration, local grid traversals nullify random behaviour and thus enhance computing accuracies.

3.4.2.MSBOA (Mutation Score Butterfly Optimization Algorithm)

MSBOA, inspired by butterfly food scavenges (Arora, S & Singh, 2015);(Arora & Singh, 2019) produces optimal feature selections from datasets. The steps followed in the algorithm are detailed below:

1. Butterflies release fragrance (score) and are attracted towards each other based on this fragrance (classification accuracy);
2. Butterflies move in a random manner towards the butterfly with highest fragrance ;
3. Butterfly's intensity is dependent on score and classifier accuracy.

In addition, MSBOA uses a switch probability p, to switch between global and intensive local searches.

3.4.3.SMFO (Score Moth-Flame Optimization)

SMFO algorithm is based on behaviour of insects, similar to butterfly behaviour. SMFO algorithm assumes features are moths and their selection is based on their positions. Moth's flight can uni-dimensional or hyper dimensional when they change their feature positions (vectors). M value of each

samples first row is passed on a fitness function whose output is assigned as the fitness value to the corresponding moth.For example, OM_1 in matrix OM. All moths are an array with their fitness values. Thus, initial solutions are generated by I and objective function values are computed (Mirjalili, 2015). This function can be used on any random distribution and is the default function of SMFO.

3.4.4.Majority Voting Strategy (MVS) for ranking by ensembles

In this study's construction of ensembles of rankers, each ensemble outputs a ranked list based on feature's relevance and resulting feature ranks are aggregated to form a single ensemble rank list (features overall scores). If L_k is the resultant ranked list of a feature selection algorithm to the k[th] sample, then for features f_i $(i = 1, \dots, N)$, the overall score is computed using equation (1),

$$score_i = score(f_i) = aggr(r_{i1}, r_{i2}, \dots . r_{iB}) \quad (1)$$

Where, r_{ik} -i[th] feature's rank in k[th] rank list and aggr – aggregating function. The resultant overall scores is then used to arrange the ordered features in an ascending order in the final ensemble output rank where a threshold value can be used as a cut off point to generate highly discriminative feature subsets. The generation of rank lists three steps are used to generate reports by the three individual feature selectors. MVS is a simple technique for combining classifiers. The first step of MVS is based on consensus amongst feature selector's results. The decision criteria for a feature to be in the list requires the feature being selected by a minimum of 2 of the 3 feature selectors or when there is a unique consensus amongst the 3 feature selectors for the same feature (Fatih et al., 2019).

3.5. Proposed WSMOTE Algorithm

WSMOTE algorithm is used to oversample minority instances. The algorithm assigns weights to minority data samples in its generation of synthetic data. This work uses WSMOTE to normalize imbalances in datasets. WSREMOTE increases minority instances in way that the classes are almost equal in percentage as it has lesser issues in accurate classifications. Parameters namely T (Minority class samples), N (Percentage of required oversamples) and k (the initial value of clustering minority samples) passed initial values. WSMOTE's oversampling is based on SMOTE's method of weight assignments to minority data samples (Prusty et al., 2017) where WSMOTE uses Euclidean distances of minority data samples from other minority data samples resulting in a weight matrix depicted in figure 2.

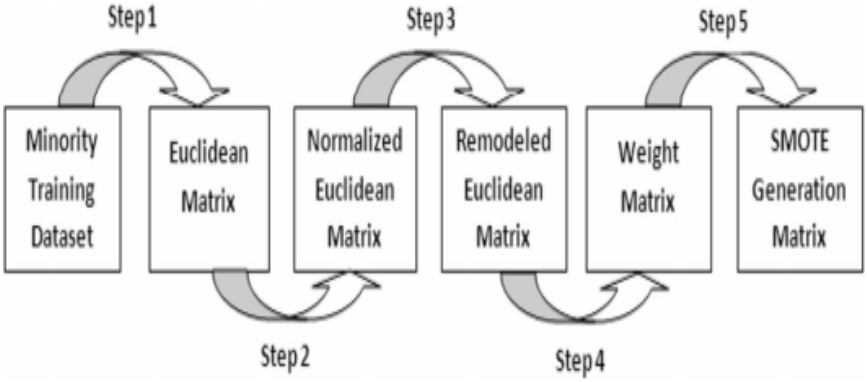

Figure 2 . WSMOTE Block Diagram

The resultant weight matrix and percentage of synthetic data generated generates a SMOTE matrix based on equation (2) which indicates the synthetic data count for generation of specific minority data samples.

$$[\text{SMOTE Generation Matrix}]_{T\times 1} = \frac{N \times T}{100} [\text{Weight Matrix}]_{T\times 1} \quad (2)$$

WSMOTE Steps:

1. Assuming the dataset has T minority samples, each with C number of features, then Euclidean distances of T samples from each other are computed using equation (3).

$$\mathrm{ED_i}(\mathrm{m_i}, \mathrm{m_j}) = \sqrt{\sum_{\mathrm{k=1}}^{\mathrm{C}} \left(\mathrm{m_{i,k}} - \mathrm{m_{j,k}}\right)^2} \quad (3)$$

Where, $\mathrm{ED_i}(\mathrm{m_i}, \mathrm{m_j})$ - Euclidean distance between i^{th} and j^{th} samples, k - k^{th} samples's attribute, $\mathrm{i} \in [1, 2, \dots, \mathrm{T}]$, $\mathrm{j} \in [1, 2, \dots, \mathrm{T}]$ and $\mathrm{j} \neq \mathrm{i}$. The total of each j^{th} minority sample is given by $\mathrm{ED_i}$ while The total of all minority Euclidean Distances is given by $\mathrm{ED} = [\mathrm{ED_1}, \mathrm{ED_2}, \dots, \mathrm{ED_T}]'$.

2. The obtained matrix ED matrix is regularized using $\mathrm{ED_{max}}$ (maximum Euclidean Distance) and $\mathrm{ED_{min}}$ ((minimum Euclidean Distance) which results in NED (Normalized Euclidean Distance) matrix given by equation (4). The normalization process map numbers in the interval [0,1].

$$\mathrm{NED_i} = \frac{\mathrm{ED_i} - \mathrm{ED_{min}}}{\mathrm{ED_{max}} - \mathrm{ED_{min}}} \quad (4)$$

3. The resultant NED matrix of (3) is Remodelled called RNED matrix which implies minority data percentage defines the generation of synthetic data corresponding to N %. RNED matrix is generated by subtracting NED matrix values from the total of all NED matrix values and represented as equation (5).

$$[\mathrm{RNED}]_{\mathrm{T\times 1}} = \mathrm{Sum(NED)} - [\mathrm{NED}]_{\mathrm{T\times 1}} \quad (5)$$

The ultimate weight matrix is computed by finding the fraction of data share for a minority class with respect to the total shares in RNED matrix given by equation (6).

$$[\mathrm{Weight\ Matrix}]_{\mathrm{T\times 1}} = \frac{[\mathrm{RNED}]_{\mathrm{T\times 1}}}{\mathrm{Sum(RNED)}} \quad (6)$$

The resultant weight matrix of (5) generates SMOTE matrix of equation (1).

3.6. Ensemble Classifier

ECs (Ensemble Classifiers) are powerful class of MLTs as they combine predictions of multiple classifiers. n samples of HCC data are generated in this work for training classifiers on HCC data . Further, the use of ECs in this work improved predictions by combining KSVM and FCNN classifier outputs. This approach contributed to better predictive performances when compared to single predictive models. The base of ECs is learning with a set of classifiers is allowing a voting scheme for the classifiers. Bagging classifier outputs are aggregated for selecting the best prediction. Different classifiers specialize in distinct feature spaces, which enables bagging predictions of every model to reach an ultimate prediction.

3.6.1.KSVM Classification

SVMs are used for two class classification (Zanaty, 2012). In its training of HCC data, SVMs mark each instance into one of the two categories i.e. SVM models predict if a new example can be categorized into a class or not. SVMs consider samples as data points in a space and map them into separate categories with a clear line of demarcation called the hyperplane. Data samples are thus mapped into one category or the other based on the hyperplane. KSVM creates non-linear data by applying kernel tricks to maximize the hyper plane (Hsieh et al., 2014), but with the use of dot product instead of a non-linear kernel function. Thus, allowing maximum-margin hyperplane fits in a transformed feature space. KVSM transformations may be non-linear in a high dimensional feature space. Thus, KVSM used in this work generates hyperplanes from non-linear data in high-dimensional feature spaces Kernel formulation in SVMs with support vectors $z_1, z_2, \dots, z_N$ and weights $w_1, w_2, \dots, w_N$ can be depicted as equation (7),

$$\mathrm{F(x)} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{w_i k(z_i, x)} + \mathrm{b} \quad (7)$$

Kernel tricks are powerful tools as they bridge linearity to non-linearity for algorithms that depend on the dot product between vectors. Kernel function usage can be used to transform data points into higher-dimensional space without explicit input mappings. It is clear that the linear kernel $k(x, y) = x^T y + c$ is the simplest kernel function given by a common inner product $< x, y >$ with an optional

constant **c.** The proposed . KSVM, inherited and extended from SVM, uses Kernel property and overridden computations to include chosen Kernel computations in the model.

3.6.2.Fuzzy Convolutional Neural Network (FCNN) Classification

This work uses FCNN for classifying the dataset. This works, FCNN uses 4 layers (input, convolution, pooling, and soft max) for ensuring accuracy of predictions with reduced computations. It uses two of CNN's layers namely input and convolution.

(i) Input Layer: This layer is trained with N × k neurons, where k is input data's variate number and N is the data length.

(ii) Convolution Layer: This layer convoluted data input from the preceding layer using convolution filters (Williams & Li, 2018).

(iii) Pooling Layer: This layer down samples or minimizes parameter count or reduces dimensionality. It is a masking operation with a sliding window on the input matrices where it moves with the size of the convolution kernel and only a single calculation is executed.

(iv) Softmax/Fully Connected Layer: Activation functions produce nonlinear outputs by combining linear networks A Softmax function is a squashing function in which the layers determine multi-class probabilities.

(v) Output Layer: In neurons/nodes corresponding to n features classes are output in this layer where they are fully connected to the feature layer. The maximum output neuron is treated as class labels from inputs in classifications.

The Bootstrap Aggregation approach incorporates the predictions from several machine learning algorithms to produce more accurate predictions than any single model. It's an ensemble meta-learning technique that uses different partitions of the training data to train classifiers like KSVM and FCNNs and the final prediction for the input vector is formed by combining the predictions of all classifiers. As a result, the bagging technique's final effect is to reduce model variance, making the prediction process more noise independent.

Algorithm 1. Bagging Technique

Input : Total N training samples

Output: Classified results

1. N training subsets are created with
   - $p \leq n$ bootstrap data samples drawn with or without replacement from the training set of n data samples
   - $q \leq m$ attributes from the original m dimensions of the dataset
2. A prediction model $h_T(x)$ is trained on each bootstrap training subset $T = 1, \ldots N$. This leads to the final ensemble model $H(x) = \{h_T(x), T = 1, .. N\}$
3. To apply the ensemble model H(x), all predictions models $h_T(x)$ are run on the input data sample x
4. To final the prediction of the ensemble $H(x)$ is based on a combination of the predictions produced by all models $h_T(x)$

In a classification problem, the majority vote or the average class score can be used. The majority vote takes the class predicted by classifiers $h_T(x, c_i)$ as the final class. The average class score takes the class predicted by the highest average score calculated on all classifiers $h_T(x, c_i)$:

$$c_i = \arg\max_i \left\{ \frac{1}{N} \sum_{T=1}^{N} h_T(x, c_i) \right\} \quad (8)$$

In numerical prediction problem, the average value calculated on all models $h_T(x)$

$$y_p = \frac{1}{N} \sum_{T=1}^{N} h_T(x) \quad (9)$$

4. Results And Discussion

This section displays this work's numerical evaluation of methods using various performance metrics. MATLAB R 2016a was used for simulations of the proposed and existing techniques and benchmarked with three datasets detailed below.

Dataset 1: Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors. This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90". The size of the dataset is 22.8 KB.

Dataset 2: The actual dataset is from University of California, Irvine (UCI) machine learning repository webpage. The dataset consists of 165 patients and 50 attributes for diagnosed with Hepatocellular Carcinoma (HCC), and includes both missing values and imbalance nature of class label. In this study, the HCC survival dataset missing values were imputed by Improved Fuzzy C Means (IFCM).

Dataset 3: The gene expression dataset is collected from CancerLivER. Three gene expressions such as GSE102079, GSE107170, and GSE25097 have been used for implementation. GSE102079 includes of 257 patients with 22048 gene expression microarray (Robust Multi Array (RMA)) for Hepatocellular Carcinoma (HCC), GSE107170 includes of 307 patients with 22048 gene expression microarray(RMA) for HCC, GSE25097 includes of 557 patients with 22048 gene expression microarray(RMA) for HCC(See Table 1). https://webs.iiitd.edu.in/raghava/cancerliver/browse_sub1.php?token=Hepatocellular&col=12 is used for dataset collection.

4.1. Performance Measures

The performance measures considered in this work are listed as follows: Precision, Recall, F-measure and Accuracy.

4.2. Precision

Proportion of correctly classified positive samples to total count of positive predictions on samples given by equation (10),

Precision = True Positive/ (True Positive + False Positive) = TP/ (TP + FP) (10)

4.3. Recall

Proportion of correctly classified samples positive to the total count of positive samples given by equation (11),

Recall = True Positive/ (False Negative+True Positive) = TP / (FN + TP) (11)

4.4. F-Measure

It is the harmonic mean of precision and recall, also called F-measure and computed using equation (12),

F-Measure = 2* (Precision * Recall) / (Precision + Recall) (12)

4.5. Accuracy

Ratio of correctly classified samples to the total samples count given by equation (13).

Accuracy = TP + TN/ (TP + TN + FP + FN) (13)

4.6. Results Comparison

The performance comparison results of the proposed EC, existing Modified Convolutional Neural Network (MCNN) and FCNNs on three datasets in terms of performance metric values are given in the following table 2.

Table 2. Performance Evaluation Results Vs. Methods

Metrics	Liver Disease - Methods			HCC - Methods			Gene expression data – methods		
	MCNN	FCNN	Ensemble classifier	MCNN	FCNN	Ensemble classifier	MCNN	FCNN	Ensemble classifier
Precision (%)	88.57	90.78	91.6667	74.3889	89.2308	92.3077	66.1326	71.2418	94.0768
Recall (%)	94.11	97.36	99.8996	61.0499	85.4169	98.6784	82.2747	91.5966	95.7265
F-Measure(%)	91.25	93.96	95.4043	67.8635	85.4289	95.1637	73.3258	80.1471	94.6131
Accuracy (%)	90.75	92.48	99.8012	82.8283	85.8586	97.6923	88.0399	88.3268	96.5517
Time Seconds)	2.9931	2.5479	2.0333	2.4384	2.0825	2.07921	77.4492	51.0988	8.619

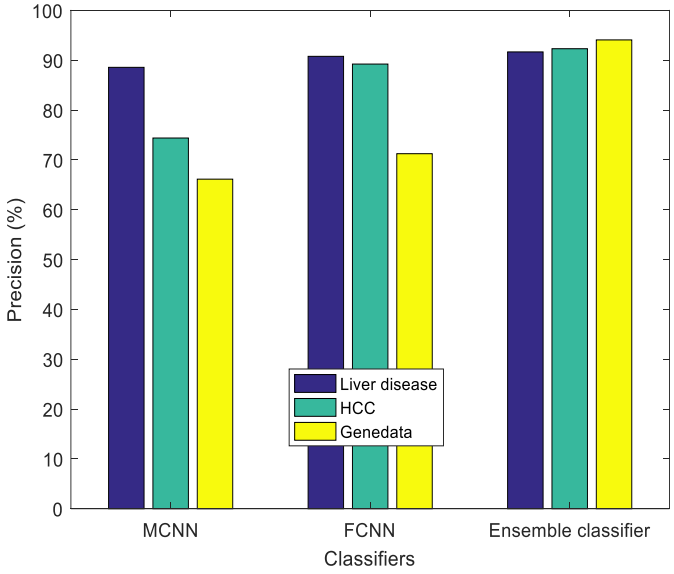

Figure 3. Precision Results Comparison Vs. Methods

Figure 3 shows the precision results comparison of three classifiers such as MCNN, FCNN and proposed ensemble classifier to three datasets. The proposed ensemble classifier gives higher precision results of 91.6667%, whereas other methods such as MCNN and FCNN gives the precision results of 88.57% and 90.78% respectively for liver disease dataset (See Table 2). This significant performance of proposed system is due to the ensemble of classifiers.

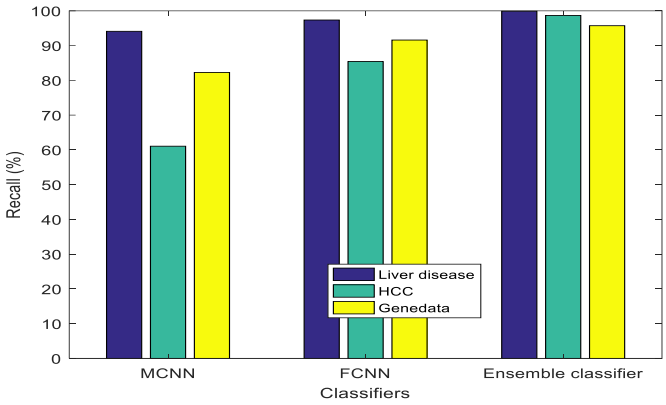

Figure 4. Recall Results Comparison Vs. Methods

Figure 4 shows the recall metric with respect to three different datasets such as liver disease, HCC and gene data of three classifiers such as MCNN, FCNN and proposed ensemble classifier. From the figure 4 it concludes that the proposed FCNN classifier gives higher recall results of 99.8996%, whereas other methods such as MCNN and FCNN gives the recall results of 94.11% and 97.36% respectively for liver disease dataset (See table 2).

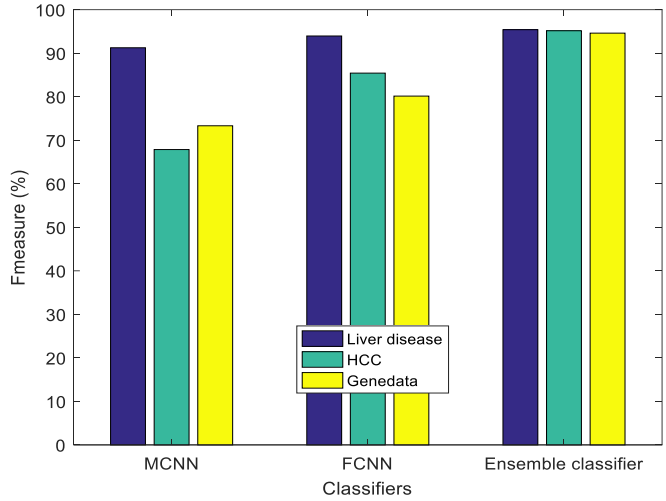

Figure 5. F-Measure Results Comparison Vs. Methods

Figure 5 shows the f-measure results comparison of three different datasets such as liver disease, HCC and gene data with respect to classifiers such as MCNN, FCNN and proposed ensemble classifier. The proposed ensemble classifier gives of 95.4043%, whereas other methods such as MCNN and FCNN gives the f-measure results of 91.25% and 93.96% respectively for liver disease dataset (See Table 2). Thus, the performance of the proposed ensemble classifier is efficient and better when compared to the existing model.

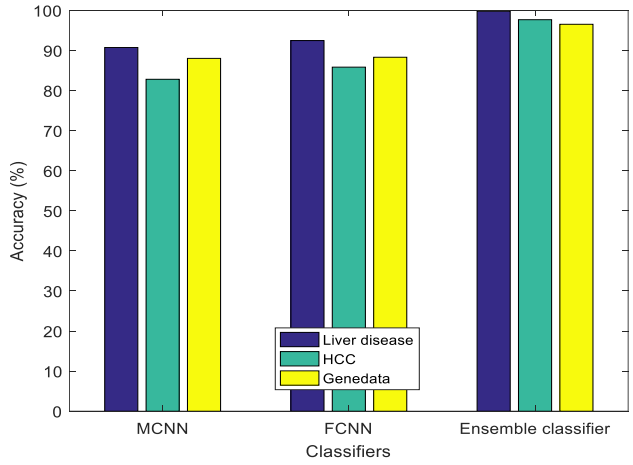

Figure 6. Accuracy Results Comparison Vs. Methods

Accuracy results comparison with respect to three classifiers in three datasets is shown in the figure 6. From the results it concludes that the proposed ensemble classifier gives higher accuracy results of 99.8012% for liver disease dataset, whereas other methods such as MCNN and FCNN gives the accuracy results of 90.75% and 92.48% respectively(See Table 2). From this analysis it is proved that the proposed shows better performance than the existing technique. Proposed ensemble classifier, imbalanced dataset issue is solved before classification which may improves the accuracy of the classifier than existing classifiers.

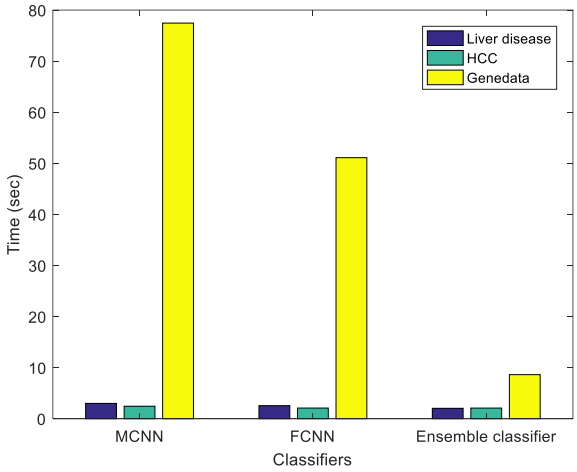

Figure 7. Time Results Comparison Vs. Methods

Figure 7 shows the time comparison results of three different datasets with three classifiers such as MCNN, FCNN and Ensemble classifier. From the figure 7 it concludes that the proposed ensemble classifier has takes lesser time of 2.0333 seconds, whereas other methods such as MCNN and FCNNs has takes more time of 2.9931 seconds and 2.5479 seconds respectively for liver disease dataset (See table 2).

5. Conclusion and Future Work

Owing to its rapid development, HCC is a malignant disease with few therapeutic options. Assess the risk of HCC in various groups of average or high risk. The aim of this study was to create and validate a new ensemble classifier-based risk prediction model for HCC growth, as well as to compare it to previously published risk models. The WSMOTE algorithm is implemented to address the problem of dataset imbalance. The WSMOTE method is an oversampling method that assigns weights to determine the number of new synthetic data that must be produced for each minority data sample using SMOTE. Values that are missing are substituted using IFCM clustering and thus enhances analysis accuracy. The imputed features are then reduced by selecting required feature using IEFS algorithm is performed based SAFSA, MSBOA and SMFO. Ensemble classification is performed by combining the results of two classifiers namely KSVM and FCNNs. The proposed scheme has been evaluated with MATLAB R 2016a simulations. The experimentation results of the proposed research work demonstrate better performances when judged on the metrics of Precision, Recall, F-measure, and accuracy. In the future work, other classifiers will be implemented to increase the prediction rate of the system.

References

1. Indhumathy, M., Nabhan, A. R., & Arumugam, S. (2018). A weighted association rule mining method for predicting HCV-human protein interactions. *Current Bioinformatics*, *13*(1), 73-84.
2. Villanueva, A. (2019). Hepatocellular carcinoma. *N. Engl. J. Med.* 380, 1450–1462.
3. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *68*(6), 394-424.
4. Kanwal, F., & Singal, A. G. (2019). Surveillance for hepatocellular carcinoma: current best practice and future direction. *Gastroenterology*, *157*(1), 54-64.
5. Zhou, M., Zhao, H., Xu, W., Bao, S., Cheng, L., & Sun, J. (2017). Discovery and validation of immune-associated long non-coding RNA biomarkers associated with clinically molecular subtype and prognosis in diffuse large B cell lymphoma. *Molecular cancer*, *16*(1), 1-13.

6. Qu, K., Gao, F., Guo, F., & Zou, Q. (2019). Taxonomy dimension reduction for colorectal cancer prediction. *Computational biology and chemistry*, *83*, 1-7.
7. Guan, Q., Chen, R., Yan, H., Cai, H., Guo, Y., Li, M., & Guo, Z. (2016). Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. *Oncotarget*, *7*(42), 68909–68920.
8. Shimizu, T., Nemoto, T., & Tokuda, Y. (2018). Effectiveness of a clinical knowledge support system for reducing diagnostic errors in outpatient care in Japan: A retrospective study. *International journal of medical informatics*, 109, 1-4.
9. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, *250*, 113-141.
10. Ali-Gombe, A., & Elyan, E. (2019). MFC-GAN: class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, *361*, 212-221.
11. Stefanowski, J. (2016). Dealing with data difficulty factors while learning from imbalanced data. In *Challenges in computational statistics and data mining*, pp. 333-363.
12. Chen, K. H., Wang, H. W., & Liu, C. M. (2020). Applying Artificial Intelligence to Survival Prediction of Hepatocellular Carcinoma Patients. In *Proceedings of the 2020 4th International Conference on Deep Learning Technologies (ICDLT)*, pp. 135-139.
13. Dong, R. Z., Yang, X., Zhang, X. Y., Gao, P. T., Ke, A. W., Sun, H. C., ... & Shi, G. M. (2019). Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. *Journal of cellular and molecular medicine*, *23*(5), 3369-3374.
14. Sato, M., Morimoto, K., Kajihara, S., Tateishi, R., Shiina, S., Koike, K., & Yatomi, Y. (2019). Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma. *Scientific reports*, *9*(1), 1-7.
15. Gui, T., Dong, X., Li, R., Li, Y., & Wang, Z. (2015). Identification of hepatocellular carcinoma–related genes with a machine learning and network analysis. *Journal of Computational Biology*, *22*(1), 63-71.
16. Iwahashi, S., Ghaibeh, A. A., Shimada, M., Morine, Y., Imura, S., Ikemoto, T., & Hirose, J. (2020). Predictability of postoperative recurrence on hepatocellular carcinoma through data mining method. *Molecular and Clinical Oncology*, *13*(5), 1-1.
17. Cao, Y., Fan, J., Cao, H., Chen, Y., Li, J., Li, J., & Zhang, S. (2021). Prediction model for recurrence of hepatocellular carcinoma after resection by using neighbor2vec based algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(2), 1-13.
18. Chiu, H. C., Ho, T. W., Lee, K. T., Chen, H. Y., & Ho, W. H. (2013). Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network. *The Scientific World Journal*, pp.1-10.
19. Zhang, Z. M., Tan, J. X., Wang, F., Dao, F. Y., Zhang, Z. Y., & Lin, H. (2020). Early diagnosis of hepatocellular carcinoma using machine learning method. *Frontiers in bioengineering and biotechnology*, pp. 1-9.
20. Hon, J. S., Shi, Z. Y., Cheng, C. Y., & Li, Z. Y. (2020). Applying Data Mining to Investigate Cancer Risk in Patients with Pyogenic Liver Abscess. In *Healthcare*, 8 (2), 1-15.
21. Książek, W., Abdar, M., Acharya, U. R., & Pławiak, P. (2019). A novel machine learning approach for early detection of hepatocellular carcinoma patients. *Cognitive Systems Research*, *54*, 116-127.
22. Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, *58*, 49-59.
23. Divya, R., & Radha, P. (2019). An Optimized HCC recurrence prediction using APO algorithm multiple time series clinical liver cancer dataset. *Journal of medical systems*, *43*(7), 1-12.
24. Huang, J., Mao, B., Bai, Y., Zhang, T., & Miao, C. (2020). An Integrated Fuzzy C-Means Method for Missing Data Imputation Using Taxi GPS Data. *Sensors*, *20*(7), 1992.
25. Arora, S., & Singh, S. (2015). Butterfly algorithm with levy flights for global optimization. In *2015 International conference on signal processing, computing and control (ISPCC)*, pp. 220-224.

26. Arora, S., & Singh, S. (2019). Butterfly optimization algorithm: a novel approach for global optimization. *Soft Computing*, *23*(3), 715-734.
27. Mirjalili, S. (2015). Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-based systems*, *89*, 228-249.
28. Prusty, M. R., Jayanthi, T., & Velusamy, K. (2017). Weighted-SMOTE: A modification to SMOTE for event classification in sodium cooled fast reactors. *Progress in nuclear energy*, *100*, 355-364.
29. Zanaty, E. A. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, *13*(3), 177-183.
30. Hsieh, C. J., Si, S., & Dhillon, I. (2014). A divide-and-conquer solver for kernel support vector machines. In *International conference on machine learning*, pp. 566-574.
31. Fatih, A. B. U. T., AKAY, M. F., & GEORGE, J. (2019). A robust ensemble feature selector based on rank aggregation for developing new VO\textsubscript {2} max prediction models using support vector machines. *Turkish Journal of Electrical Engineering & Computer Sciences*, *27*(5), 3648-3664.
32. Williams, T., & Li, R. (2018). Wavelet pooling for convolutional neural networks. In *International Conference on Learning Representations*, pp. 1-12.