Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 4759 - 4782
Received 05 March 2021; Accepted 01 April 2021.

Intelligence Ensemble-Based Feature Selection (Iefs) Algorithm and Fuzzy Convolutional Neural Network (Fcnn) for Hepatocellular Carcinoma (Hcc) in Liver Disease System

Jeyalakshmi ,Rangaraj R

[1]Associate Professor ,Department of Computer science ,Hindhusthan College of Arts and Science, Coimbatore ,Tamilnadu.

[2]Professor and Head, PG and research, Department of Computer science ,Hindhusthan College of Arts and Science Coimbatore, Tamilnadu.

ABSTRACT: Predicting HCC (Hepato Cellular Carcinoma) risks is an important pre-disposing condition for patients with affected with fatty liver diseases who are non-alcoholic and patients recovering from hepatitis C viral infections. Though HCC can be diagnosed early, missing values and innumerable features in patient's data/datasets makes it a complicated issue. Missing values can be neutralized by ignoring or removing or imputing them where ignoring or removing information reduces processing information while accuracy while analyzing them. This research work imputes missing values using an Improved FCM (Fuzzy C Means) clustering and thus enhances analysis accuracy. Improved FCM is used as data can belong to multiple groups based on their membership values. The imputed features are then reduced by selecting required feature using an intelligent ensemble based feature selection. Multiple ensemble methods are used in the study for selecting the optimal feature subset which includes SAFSA (Score based Artificial Fish Swarm Algorithm), MSBOA (Mutation Score Butterfly Optimization Algorithm) and SMFO (Score Moth-Flame Optimization). These ensemble methods result in enhancing prediction accuracy in classifications executed using FCNNs (Fuzzy Convolution Neural Networks). The proposed scheme when tested on MATLAB (Matrix Laboratory) could classify better than most other methods as it had better values in terms of precision, recall, F-measure and accuracy.

INDEX TERMS: Liver disease prediction, Hepatocellular Carcinoma (HCC), missing data imputation, clustering, Dimensionality reduction, Feature Selection (FS), Intelligence Ensemble-Based Feature Selection (IEFS), and Fuzzy Convolutional Neural Network (FCNN).

1. INTRODUCTION

Liver is 2nd most important organ for humans. Liver plays a significant role in human metabolism where one vital function is the decomposition of RBCs (Red Blood Cells) [1]. LCs (Liver Cancers) are aggressive leading to malignant neoplasms. The past decades have drastically improved incidences of LCs on a global scale including developed countries. HCC is a major histological cancer found in the liver causing increased mortality rates in the US and ranks 4th in global cancer deaths [2]. HCCs expand into CLDs (Chronic Liver Diseases) when infected by HBVs (Hepatitis B Virus (HBV) or HCVs (Hepatitis C Virus) or NAFLDs (Non-Alcoholic Fatty Liver Diseases) or alcoholic abuses [3]. These etiological factors encourage progression of liver fibrosis finally resulting in cirrhosis which account to around 2% or 1.32 million perennial deaths [3,4]. Though effective anti-viral medicines have been developed for hepatitis viruses, they have not managed to reduce risks of HCC and more specifically in the case of liver fibrosis [5, 6]. Arrest of HCC progression has been ineffective in NAFLDs and alcoholically abused patients [3]. HCV infected patients show unsynchronized symptoms making HCC risk predictions complex. The risk factor changes dramatically with the development of cirrhosis and eliminating HCV. Further, HCC risk factors also abruptly change with patient's age or hypertension increases or increases in liver stiffness or decline in platelet counts. Analysis of controlled therapy trials of IFN (Interferon) showed decrease in HCC developments in liver cirrhosis type C patients [7].

DMTs (Data Mining Techniques) are one option to identify diseases. MLTs (Machine Learning Techniques), a part of DMTs combine the fields of mathematics and computer science in their implementation of algorithms. DMT algorithms maximize predictions using analytic or probabilistic models on both static and dynamic data sources [8]. MLTs widely used in cancer researches include ANNs (Artificial Neural Networks), NB (Naive Bayes), SVMs (Support Vector Machines), and DTs (Decision Trees). These techniques use inputs from clinical, image and gene data for analyzing and developing effective and accurate prediction models. MLTs have also been used systems for early diagnosis of cancers as they have the ability to recognize related patterns in patient's data or assess risk factors from complex patient datasets and thus use this information effectively to predict future trends of patients affected with cancer [9,10]. Multitude of studies has proved that MLTs can significantly help in identifying cancers, but missing values in information may lead to biased predictions [11]. Further, cancer predictions using MLTs are less biased when imputation methods are used on the input data [12,13]. Imputation methods can be based on mean/regression values or FCM clustering outputs [14].

Feature subset selections help in enhancing MLT model performances. Prior studies focused on finding suitable subset of features for constructing inference models. Selection of optimal features from an input dataset is one problematic area for MLT; however, current "ensemble" methods which combine outputs from model sets have greatly improved generalization accuracy of models. Ensemble models can search voluminous feature spaces non-trivially [16] which can be solved by using intelligence methods [15]. The rationale of this study is to detect HCC in patients with liver diseases using DMTs. This research work uses data mining steps of data imputations, dimensionality reductions, feature selections, classifications and result evaluations.

2. LITERATURE REVIEW

A scoring system for HCC identification was proposed by Ma et al [17] in their study. Their predictive scoring system assessed HCC development stages and suggested clinical/public health improvement approaches. The scheme's risk assessment was both cost and effort effective and suitable for personal surveillance. The study also summarized HCC risk prediction features in varied population and perspectives.

Five years of patients suffering from Chronic hepatitis C was analyzed in the study by Kurosaki et al [18]. Their analysis was based on a simple model that identified high risks of HCC development. Their proposed DLt was a predictive model which was validated using thousand seventy two patients where four hundred and seventy two had SVRs (Sustained Virological Responses) while six hundred patients were in the non-SVR category. The predictions were modelled on available disease features for identifying HCC. The model used available medical test data for inferences and was easy to implement in clinics.

The study by Cao et al [19] predicted HCC recurrences using neighbor2vec algorithm. Their scheme operated in three phases to detect HCC occurances. Pearson correlation coefficient explored independent variables in the first phase. KNN (K-Nearest Neighbors) generated a K-vectors list (neighbor2vec) for each patient when resulting correlations were low. The generated vectors lists were processed by MLTs including LRs (Logistic Regressions), KNN, DTs, NB and DNNs (Deep Neural Networks) for neighbor2vec based predictions. The scheme's experimental results on

Shandong Provincial Hospital data in China showed the model's superior performance over the other models. The proposed model when compared with NB achieved 83.02% accuracy, 82.86% recall, and 77.6% precision.

Swarm Optimization technique was used by Demir et al [20] in their study. The study filled missing data features using statistical averages and used swarm optimization for feature selections. The study's data taken from a HCC survey had forty nine features. CDO (Chaotic Darcy Optimization) technique selected features (31) from the HCC dataset. Their results were satisfactory.

MLTs were used by Książek et al [21] in their study to detect HCC from 165 patient's data. The data was pre-processed using normalization followed by GA (Genetic Algorithm) based processing with a fivefold stratified cross-validation both for parameter optimization and feature selection. The selected features were then fed into SVMs using a two level genetic optimizer (genetic training). The study's feature selections showed scores in terms of high accuracy (0.8849) and F1-Score (0.8762). The scheme was found suitable when tested on huge database and in aid of clinicians.

RFEs (Recursive Feature Eliminations) combined with SVMs were proposed in the study by Dong et al [22]. The study analyzed methylation chip data encompassing 50 normal and 377 HCC samples from the TCGA database. The scheme screened 47,099 samples from 134 methylated sites using SVM-RFE, Cox regression and FW-SVM algorithms. Their model projected patient survivals based on high, intermediate and low risk categories assessed by the model. The model's 10-fold cross-validation of 0.95 showed it had good predictive power as it classified 26 out of 33 samples.

Expression Orderings were used by Zhang et al [23] in their study. Their approach was designed on microarray data encompassing 242 C2 HCC and 1091 HCC samples. The scheme's REOs (Relative Expression Orderings) extracted numerical descriptors from gene expression profiles datasets. The dataset's unrelated features were eliminated using mRMR (maximum Redundancy Minimum Relevance) and incremental feature selections. The scheme produced 11 gene pairs to discriminate features for HCC recognition and was tested on many datasets. This discovered 11 gene proved that they could be used as signatures for identifying HCC and its surrounding non-cancerous tissues in C2HCC patient's biopsy or even inaccurately sampled specimens.

DTs were used to predict HCC in the study of Omaran et al [24]. The study used DTs as they are reliable, economical and easy to interpret. There were 29 features from which AFP (Alpha FetoProtein) values greater than or equal to 50.3ng/ml and in addition Sex, cirrhosis traces, AST>64U/L, and as cites were taken as the base for predicting HCC. The study obtained recall (sensitivity) values of 83.5% and precession (specificity) values of 83.3% on regular patient's data where rightly classified instances were 259 (82.2%) while 56 (17.8%) were not classified properly. The use of AFP values for predictions was the novelty of the scheme which predicted more than two factors for successful predictions and achieved 82% specificity and 96% sensitivity

Multitude of MLTs was used by Tian et al [25] in their study. Their study's model was based on XGBoost (extreme Gradient Boosting), RF (Random Forest), DTs and LR. The scheme's optimality was assessed using AUCs (Area Under the receiver operating characteristic Curves). The feature importance plot of XGBoost displayed the level of HBsAg and age which were important variables for tracing HBV DNA. Their classification results of HBVs demonstrated the potential of MLTs in identifying HBsAg (Hepatitis B surface antigen) that could ber used by clinicians.

3. PROPOSED METHODOLOGY

This proposed method is a mix of multiple techniques. Dataset's missing information is imputed; IFCM (Improved Fuzzy C Means) clustering replaces missing values. An Intelligent Ensemble

Received 05 March 2021; Accepted 01 April 2021.

Feature Selection algorithm based on several methods and voting function is used. HCC prediction is executed using FCNNs (Fuzzy Convolution Neural Networks). The flow of this research work is depicted as Figure 1.

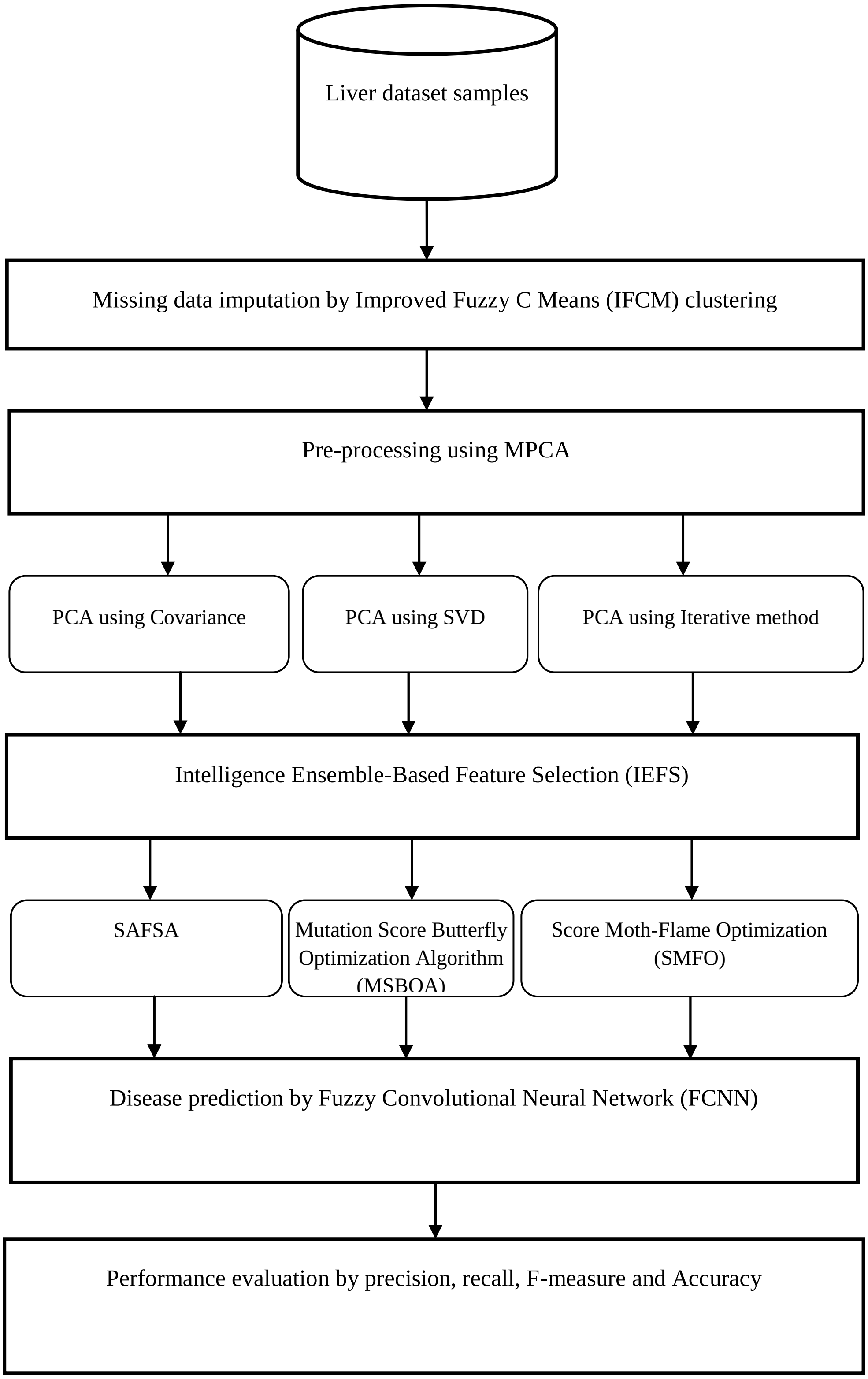
Liver dataset samples
Missing data imputation by Improved Fuzzy C Means (IFCM) clustering
Pre-processing using MPCA
PCA using Covariance
PCA using SVD
PCA using Iterative method
Intelligence Ensemble-Based Feature Selection (IEFS)
SAFSA
Mutation Score Butterfly Optimization Algorithm (MSBOA)
Score Moth-Flame Optimization (SMFO)
Disease prediction by Fuzzy Convolutional Neural Network (FCNN)
Performance evaluation by precision, recall, F-measure and Accuracy

FIGURE 1. OVERALL FLOW OF THE PROPOSED SYSTEM

Received 05 March 2021; Accepted 01 April 2021.

3.1. DATASET DESCRIPTION

Three benchmark datasets from the UCI's (University of California, Irvine) ML (Machine Learning) repository webpage [26], 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 was framed from a Portugal University Hospital encompassing risk, demographic, laboratory and survival features from https://archive.ics.uci.edu/ml/datasets/HCC+Survival. It has 49 currently used HCC survivalfeatures based on EASL-EORTC (European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer) Clinical Practice Guidelines. The features can be split into quantitative (23) and qualitative (26) variables. 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 shows the dataset 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. The attribute descriptions are shown in Table 2.

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. The attribute descriptions are shown in Table 3.

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)- Hepatocellular Carcinoma (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

TABLE 2. DESCRIPTION OF ATTRIBUTES AND THEIR CODES OF INDIAN LIVER PATIENT RECORDS (ILPD) DATASET

Age	Age of the patients	Yearly numeric values are in the interval [4-90]
Gender	Sex of the patients	Nominal values (Male/Female)
Total_Bilirubin (TB)	Total Billirubin in mg/dL	Numbers in the interval [0.4-75]
Direct_Bilirubin(DB)	Conjugated Billirubin in mg/dL	Numbers in the interval [0.1-19.7]
Alkaline_Phosphotase	ALP in IU/L	Numbers in the interval [63-2110]
Alamine_Aminotransferase	ALT in IU/L	Numbers in the interval [10-2000]
Aspartate_Aminotransferase	AST in IU/L	Numbers in the interval [10-4929]
Total_Protiens	Total Proteins g/dL	Numeric value having range [2.7-9.6]
Albumin	Albumin in g/dL	Numbers in the interval [0.9-5.5]
Albumin_and_Globulin_Ratio	A/G ratio	Numbers in the interval [0.3-2.8]
Class	Having the class value “1”represents Liver Disease Present and “2” represent Liver	Numbers 1,2

TABLE 3. DESCRIPTION OF ATTRIBUTES AND THEIR CODES OF HCC DATASET

Description: Nominal	**Code**	**Description: Continuous**	**Code**
Gender	Gender	**Alchol in gm/Day**	Grams_Day
Symptoms	Symptoms	**cigarets Packs/Year**	Packs_year
Alcohol	Alcohol	**International Normalized Ratio**	INR
Hepatitis B Surface Antigen	HBsAg	**Alpha- Fetoprotein (ng/mL)**	
Hepatitis B e Antigen	HBeAg	**Haemoglobin (g/DL)**	

Cirrhosis	Cirrhosis	**Mean Corpuscular Volume (fl)**	MCV
Endemic Countries	Endemic	**Albumin(mg/dL)**	
Smoking	Smoking	**Total Bilirubin(mg/Dl)**	Total Bil
Diabetes	Diabetes	**Alanine Transaminase(U/L)**	ALT
Obesity	Obesity	**Aspartate Transaminase(U/L)**	AST
Hemochromatosis	Hemochro	**GammeGlutamyl Transferase(U/L)**	GGT
Arterial Hypertension	AHT	**Alkaline Phosphatase(U/L)**	ALP
Chronic Renal Insufficiency	CRI	**Total Proteins(g/Dl)**	TP
Human Immunodeficiency Virus	HIV	**Number of Nodules**	Nodule
NonaIcoholicSteatohepatitis	NASH	**Creatinine(mg/dL)**	
Esophageal Varices	Varices	**Major Dimension of nodule (cm)**	Major_Dim
Splenomegaly	Spleno	**Direct Bilirubin(mg/dL)**	Dir_Bil
Portal Hypertension	PHT	**Iron (mcg/dL)**	
Portal Vein Thrombosis	PVT	**Oxygen Saturation (%)**	Sat
Liver Metastasis	Metastasis	**Ferritin(ng/mL)**	
Radiological Hallmark	Hallmark	**Description: Ordinal**	
1=Survives, 0= Died	Class	**Performance Status**	PS
Description :Integer		**Encefalopathy Degree**	Encefalopathy
Age at diagnosis	Age	**Ascites Degree**	Ascites

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 $X = \{x_1, \dots x_k\}$ and n is the count of samples. If the matrix with p features is represented by $x_k = \{x_{1k}, x_{2k}, \dots x_{jk}, \dots x_{pk}\}$ where p is the count of features the matrix structure can be depicted as Equation (1),

$$X = \begin{bmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & x_{jk} & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix} \tag{1}$$

Assuming again c is the clusters count where $1 \le i \le c$, and y_i represents a single cluster, the estimation accuracy can be increased in each y_i with p attributes and each attribute stands for the

Received 05 March 2021; Accepted 01 April 2021.

cluster's centroid, then, $y_i = \{y_{1i}, y_{2i}, \ldots y_{ji}, \ldots y_{pi}\}$ which deviates from Tang et al [28]. FCM membership degree $u(x_k, y_{ji})$, represents the distance between x_k and cluster centroid y_{ji} or $d(x_k, y_{ji})$.

$$\min J = \sum_{k=1}^{n}\sum_{i=1}^{c}\sum_{j=1}^{p} u(x_k, y_{ji})^m . d(x_k, y_{ji}) \tag{2}$$

$$d(x_k, y_{ji}) = \sum_{g=1}^{p} (x_{gk} - y_{ji})^{1/t} \tag{3}$$

$$u(x_k, y_{ji}) = \frac{1}{\left[\sum_{a=1}^{c} \left(\frac{d(x_k, y_{ji})}{d(x_k, y_{ja})}\right)^{1/(m-1)}\right]} \tag{4}$$

$$\sum_{j=1}^{p} u(x_k, y_{ji}) = 1 \tag{5}$$

$$y_{ji} = \frac{\sum_{k=1}^{n}\left[u(x_k, y_{ji})^m . x_{jk}\right]}{\sum_{k=1}^{n}\left[u(x_k, y_{ji})^m\right]} \tag{6}$$

$$x_{jk}^{*} = \sum_{i=1}^{c} u(x_k, y_{ji}). y_{ji} \tag{7}$$

$$d(x_k, y_{ji}) = \exp\left(-\gamma \left|\left|x_k - y_{ji}\right|\right|^2\right) \tag{8}$$

Where, $\left|\left|x_k - y_{ji}\right|\right|^2$ - squared Euclidean distance between any 2 data points, σ^2- free parameter and $\gamma = \frac{1}{2\sigma^2}$. The membership degrees, distances and objective functions are applied simultaneously as shown in Equation (2). The distance of a point from the cluster centroid is computed using Equation (3) where if t =1 implies it is a Manhattan distance and when equal to 2 it is the Euclidean distance [29]. Membership degree can be computed using Equation (4) and the total of membership degrees of each x_k should be 1 as shown in Equation (5). If the difference between new membership degree $u(x_k, y_{ji})^*$ and old membership degree $u(x_k, y_{ji})$ is larger than the threshold ε, the new cluster centroid should be updated by equation (6). When the optimal cluster centroids are finally obtained, the missing values can be obtained by equation (7). In the general FCM clustering algorithm, instead of considering the distance $d(x_k, y_{ji})$ by equation (3), equation (8) is replaced with kernel function.

3.3. MPCA (MODIFIED PRINCIPAL COMPONENT ANALYSIS) 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. This can be avoided by the pre-processing the input dataset. In this work, This work reduces dimensionality of features using MPCA where PCA is a multivariate dimensionality reduction method and has been used widely used to extract important features in data compressions and classifications. PCA can minimize errors and de-correlate features. In a matrix $X = [x_{ik}]_{m \times n}$, where m original samples count and n the attributes count, PCA on X can be depicted using Equation (9),

$$Y = TX \tag{9}$$

Where, T - transformed matrix, X - original matrix and Y – dimensionally reduced matrix output. T can be solved using Equation (10),

$$(\lambda I - S)U = 0 \tag{10}$$

Where, I - Unity square matrix along the diagonal, S – Original Data's covariance matrix, U - eigenvectors and λ –0020eigen values. U_j and λ_j(j =1,2,...,m) are calculated using (10) and the order of eigen values is $\lambda_1 \geq \lambda_2 \geq \ldots . \geq \lambda_m$ and $U = [U_1, U_2, \ldots., U_m] = [u_{ij}]_{m\times n}$and satisfies Equation (11), $U^T U = UU^T = I$. T is computed by inversing U. 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 recursive methods. Thus, in MPCA only samples relevant in the liver dataset get transformed (T').

$$Y' = T'X \tag{11}$$

Using (10) and (11), 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. 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. Assuming there are X fishes where$X = (d_1, d_2, \ldots, d_{mn})$, is the set of samples, then the current state of fish in relation to feature selections is X while $X_v$is the new state of the fish sample and is based on Equation (12).

$$X_v = X + Visual * r \tag{12}$$

Then the basic movement process can be expressed as in equation (13).

$$X_{next} = \frac{X_v - X}{||X_v - X||}.\mathrm{Step}.\, r + X \tag{13}$$

Where, r - random numbers between 0 and 1, , Step – a move's step size and $\mathrm{dis}(X_i, X_j)$ - distance between two fish samples [30]. In Equation (16), X is the optimum global position of food concentration and to balance convergence speed and precision, this study uses a dynamic parameter λ where (0<λ<1). The parameters adapted to this dynamic factor of SAFSA are depicted in equations (14,15,16).

$$\text{Visual} = \text{Visual}\ (1-\lambda) \tag{14}$$

$$\text{Step} = \text{Step}\ (1\text{-}\lambda) \tag{15}$$

$$\mathrm{X_{next}} = \frac{\mathrm{X_{best}} + \mathrm{X_v} - 2.\mathrm{X}}{\|\mathrm{X_{best}} + \mathrm{X_v} - 2.\mathrm{X}\|}.\mathrm{Step.r} + \mathrm{X} \tag{16}$$

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. The fitness score using Modified hardy-weinberg is depicted in Equation (17).

$$z=(p^2w_{11} + 2pqw_{12} + q^2w_{22}) \ (17)$$

Where, w_{11}- 0.1, w_{12} - 0.3, w_{22} - 0.5, p- information gain, q- entropy and z -fitness value which is maximised. SAFSA updates weight r in every iteration for convergence and thus resulting in best solution. Equations get multiplies with r.

3.4.2. MSBOA

MSBOA, inspired by butterfly food scavenges [31,32] 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..

MSBOA's perceived fragrance (f) can be defined as a function by following equation(18),

$$f = cI^a \tag{18}$$

Where, c ∈ [0, ∞] - sensory modality; I – calculated stimulus intensity using score and accuracy and associated with the objective function; and $a \in [0, 1]$ - power exponent dependent of the mutation operator or varying degree of fragrance. Two key steps of MSBOA are global and local searches where global search is movement of butterflies towards stronger fragrance (optimal selection of features) and depicted in Equation (19),

$$x_i^{t+1} = x_i^t + (r^2 * g^* - x_i^t) \times f_i \tag{19}$$

Where x_i^t - location of the ith butterfly at time t, g^* - current best location, f_i- fragrance of the ith butterfly, and $r \in [0,1]$ - random number. The local random walks is depicted as Equation (20),

$$x_i^{t+1} = x_i^t + (r^2 \times x_j^t - x_k^t) \times f_i \tag{20}$$

Where, x_j^t- position of the jth butterfly, x_k^t - position of the kth butterfly and r - random number in the interval [0, 1]. In addition, MSBOA uses a switch probability p, to switch between a global and intensive local searches.

ALGORITHM 1- MSBOA

1. **Begin**
2. Generate initial population P containing n butterflies $pop_i(i = 1, 2, \ldots, n)$,objective function $f(x), x = (x_1, x_2, \ldots, x_d)^T$, here d represents the number of dimensions.
3. Stimulus $I_i$intensity at is pop_idetermined by the fitness value f(pop_i) via the score and classification accuracy

```
4. Define sensor modality c, power exponent a by mutation operator, and switch probability p.
5. While stop criteria not met do
   5.1. For each butterfly in the population P do
           Calculate fragrance f using Equation (18).
   5.2. End for
   5.3. Evaluate and rank the population P, and find the best feature in the population.
   5.4. for each butterfly in the population P do
        5.4.1.  Generate a random number r ∼U[0, 1].
        5.4.2.  if (r < p)
                   Implement global search using Equation (19).
        5.4.3.  else
                   Implement local search using Equation (20).
        5.4.4.  end if
   5.5. end for
   5.6. Update the value of the power exponent a by mutation operator.
6. End while
7. Output the best solution and optimal value.
8. End
```

3.4.3. SMFO

SMFO 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 one-dimensional or two-dimensional or three-dimensional or even hyper dimensional by changing their feature positions (vectors). SMFO is based on population and can be depicted as Equation (21),

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & \ldots & m_{1,d} \\ m_{2,1} & m_{2,2} & \cdots & \ldots & m_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & \cdots & m_{n,d} \end{bmatrix} \tag{21}$$

Where, n - Samples count of moths and d – count of features (dimension).

Received 05 March 2021; Accepted 01 April 2021.

$$OM = \begin{bmatrix} OM_1 \\ OM_2 \\ \vdots \\ OM_n \end{bmatrix} \tag{22}$$

The first row of M for each sample is passed to the fitness function and its output assigned to the corresponding moth as its fitness value, for example, OM_1 in matrix OM. All moths are an array with their fitness values. The proposed SMFO also uses flames, similar to moths and depicted in Equation (23),

$$F = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & ... & F_{1,d} \\ F_{2,1} & F_{2,2} & \cdots & ... & F_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{n,1} & F_{n,2} & \cdots & \cdots & F_{n,d} \end{bmatrix} \tag{23}$$

It is evident from Equation (23) that M and F have equal dimensions. F stores fitness values of flames and depicted as Equation (24),

$$OF = \begin{bmatrix} OF_1 \\ OF_2 \\ \vdots \\ OF_n \end{bmatrix} \tag{24}$$

SMFO algorithm uses three functions (I, P, T) for approximating globally optimal feature selections. I generates random feature population with corresponding fitness values and the functional model is depicted by Equation (25),

$$I: \emptyset \rightarrow \{M, OM\} \tag{25}$$

P, is the main function which moves moths in the liver disease feature space. P receives M and returns its updated values using Equation (26),

$$P: M \rightarrow M \tag{26}$$

T returns the termination criterion status as true or false based on Equation (27),

$$T: M \rightarrow \{true, false\} \tag{27}$$

Thus, initial solutions are generated by I and objective function values are computed. This function can be used on any random distribution and is the default function of SMFO.

ALGORITHM 2. FITNESS CALCULATION AND GENERATE INITIAL SOLUTIONS

```
1. For $i = 1:n$
2. for $j = 1:d$
   $M(i,j) = \left(u_b(i) - l_b(i)\right)0 * rand() + l_b(i)$
3. End
4. End
5. OM=Fitness Function (M)
```

u_b - upper bound of the ithfeature l_b- lower bound of the ithfeature. On initialization, the function P runs iteratively until function T returns true. The function P moves the features around the feature search space. The position of each feature with respect to flame is updated using Equation (28),

$$M_i = S(M_i, F_j) \tag{28}$$

Where, M_i - ith moth(Feature), F$_j$ - jth flame feature, and S - spiral function. The main updates occur based on logarithmic spirals which are subject to the conditions given below:

- Spiral's initial point is a moth.
- Spiral's final point is a flame (Best features).
- Spiral's range has to be within the search space.

The logarithmic spiral of SMFO algorithm is defined as Equation (29),

$$S(M_i, F_j) = Dis_i e^{bt}.\cos(2\pi t) + F_j \tag{29}$$

Where, Dis_i- ith moth's distance for the jth flame, b - shape of the logarithmic spiral and a constant, and $t \in [1,1]$ - random number.

$$Dis_i = |F_j - M_i| \tag{30}$$

Equation (30) is used to simulate moth's spiral flying paths where its next position is defined with respect to flame flames. The parameter t defines the closeness of moth's next position with respect to flame, where t = 1 is the moth's farthest position from the flame. The search space is a hyper ellipse assuming flames are in all directions and moths move spirally as it defines how moths update their positions around flames. Hence, spiral equation allows moths to fly around flames and hence explorations and exploitations using the function is a guarantee. Figure 2 depicts logarithmic spiral and space around the flame for different values of t.

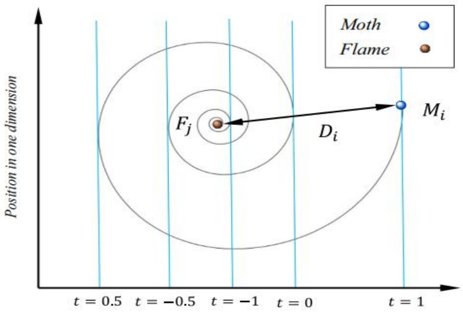

FIGURE 2. LOGARITHMIC SPIRAL, SPACE AROUND A FLAME, AND THE POSITION WITH RESPECT TO T

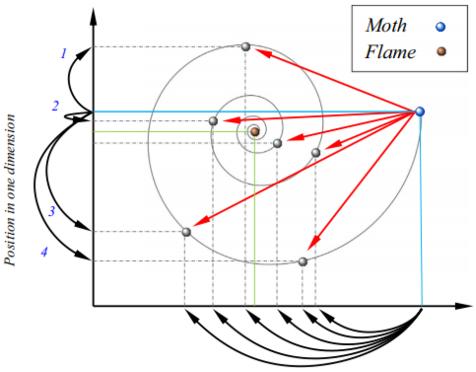

FIGURE 3. SOME OF THE POSSIBLE POSITIONS THAT CAN BE REACHED BY A MOTH WITH RESPECT TO A FLAME USING THE LOGARITHMIC SPIRAL [33]

Figure 3 depicts updates by moths around a flame conceptually where vertical axis is one dimensional represent a single variable/parameter of a given problem, but the proposed method can change all features in a problem. Dashed black lines are probable next positions of moths (blue horizontal line) around the flame (green horizontal line). The figure 3 brightly indicates a moth's one-dimensional exploration/exploitation in a search space. Explorations are next positions in the outside space between the moth and flam (Arrows 1, 3, and 4) while Exploitations are next positions lies inside the space between the moth and flame (Arrow 2).

3.4.4. MVS (Majority Voting Strategy) 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 kth sample, then for features f_i $(i = 1, \dots, N)$, the overall score is computed using Equation (31),

$$score_i = score(f_i) = aggr(r_{i1}, r_{i2}, \dots . r_{iB}) \tag{31}$$

Where, r_{ik} -ith feature's rank in kth 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. Assuming, if 3 feature selectors report different features ranks, then the second step is followed [34].

In the second step correlation scores of each feature is reported by the 3 feature selectors. Equation (32) gives the correlation score of a feature which includes computing their Inconsistency Scores (InS) across 3 feature selectors and the feature's distance from the first rank (DS). All three parameters used in the rank are depicted as equations InS (33) and DS (34) [34]:

$$CS_{fe}(ifs, i) = InS_{fe}(fs, i) + DS_{fe}(fs, i) \tag{32}$$

$$\begin{aligned} &InS_{fe}(ifs, i) \\ &= \begin{pmatrix} \left|rank_{fe}(\mathrm{MSBOA-F}) - \mathrm{rank}(\mathrm{SAFSA})\right| + |\mathrm{rank_{fe}}(\mathrm{MSBOA-F}) - \mathrm{rank_{fe}}(\mathrm{SMFO})| + \\ |\mathrm{rank_{fe}}(\mathrm{SMFO}) - \mathrm{rank_{fe}}(\mathrm{SAFSA})| \end{pmatrix} \end{aligned} \tag{33}$$

$$\begin{aligned} DS_{fe}(ifs, i) = {}& \left(rank_{fe}(\mathrm{MSBOA-F}) - 1\right) + \left(rank_{fe}(\mathrm{SAFSA} - 1)\right) \\ & + \left(rank_{fe}(\mathrm{SMFO} - 1)\right) \end{aligned} \tag{34}$$

Where, fe – feature's rank relevance generated by the function as reported by ifs (intelligent feature selector). InS shows the rank inconsistency of a feature as reported by 3 feature selectors and DS is absolute rank difference of the feature from the first rank. Lower scores imply greater relevance. Especially, a low IS implies that it has been approximated to the same rank by all 3 feature selectors. Low DS also implies it is more relevant to the 2 two features. The correlation scores calculated for each feature in a triplet candidate set is used to select features with lowest correlation values. When two features have equivalent correlations, MVS follows the third step [34]. Thus, features selected in the first or feature with highest correlations in the second step are updated to the feature rank list of the selector, assuming they may not be correct. A corrected rank moves up on the rank list of other feature selectors and the list gets rearranged and re-ranked. This procedure is repeated until the relevance of all predictor variables are reported by the feature selectors. The proposed ensemble ranker is illustrated as Algorithm 3.

ALGORITHM 3: PSEUDOCODE OF THE MVS ALGORITHM

Input: Ranking list of predictor features reported by SAFSA, MSBOA and SMFO

Output: A new MVS based ranking list of predictor features

```
1. Repeat
    1.1. Retrieve the triple of predictor features at rank i
    1.2. Apply the MVS on the triple of features
    1.3. If (Consensus for a feature is reached)
            then select the corresponding feature by majority of votes
    1.4. Else
        1.4.1. Calculate the CS for each feature in the triple candidate set by equation (32)
        1.4.2. if (there is a single minimum CS)
               Then select the feature with the lowest CS
        1.4.3. Else
               Select the predictor feature reported by the feature selector that has been assigned
               with the highest priority
        1.4.4. End if
    1.5. end if
2. Until the relevance of all predictor features has been evaluated and used for classification
```

3.5. FCNN CLASSIFICATION

This work uses FCNN for classifying the dataset. This work's GCNN 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 withN × k neurons, where k is input data's variate number andN is the data length.

(ii) Convolution layer: This layer convoluted data input from the preceding layer using convolution filters. The kernel with a moving step of 1 is depicted in Equation (35),

$$Y_{IJ} = \sum_{i=0}^{K_w} \sum_{j=0}^{K_h} x_{(I+i-1)(J+j-1)} * k_{ij} \tag{35}$$

Where, Y_{IJ} - output matrix, K_w – kernel width and K_h- kernel height, $K_w = K_h$ - kernel is a square kernel, x_{ij}- input matrix, and k_{ij} – Kernel weight computed using a fuzzy membership function depicted in Equation (36) and updated in training. A constant size after convolution is maintained by padding zeros to the edge of the input matrix.

$$fk_{ij} = f(k_{ij}, a_{ij}, b_{ij}, c_{ij}) = \left\{ \begin{matrix} 0, & k_{ij} \leq a_{ij} \\ \dfrac{k_{ij} - a_{ij}}{b_{ij} - a_{ij}}, & a_{ij} \leq k_{ij} \leq b_{ij} \\ \dfrac{c_{ij} - k_{ij}}{c_{ij} - b_{ij}} & b_{ij} \leq k_{ij} \leq c_{ij} \\ 0, & c_{ij} \leq k_{ij} \end{matrix} \right\} \tag{36}$$

Where, a, b, c - parameters of weights used for computing input weight in k_{ij}.

(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. Hence, in max pooling an N × N mask outputs 1/N times to achieve dimensionality reduction as depicted in Equation (37),

$$a_{k_{ij}} = \max_{(p,q)\in fe} (a_{kpq}) \tag{37}$$

Where, $a_{k_{ij}}$ - activation function output of the kth feature map at (i,j), a_{kpq} - input activation function at (p,q) within fe_{ij}, where $|fe|$ implies pooling region's size [35].

(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 with limitations. Hence, ReLU (Rectified Linear Unit) defined as Equation (38) is used as the activation function

$$f(x) = \begin{cases} 0, if\ x < 0 \\ x, ifx \geq 0 \end{cases} \tag{38}$$

Where, if x < 0 implies output is 0 and when x > 0 the output is x.

(v) Output layer.n 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 are treated as class labels from inputs in classifications.

4. RESULTS AND DISCUSSION

In this section, numerical evaluation of the methods is done in terms of various performance measures. MATLABR2016a simulation environment is used to implement the proposed and existing methods. The following three datasets have been used for implementation.

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.

4.1.1. Precision

Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by equation (39),

Precision = True Positive/ (True Positive + False Positive) = TP/ (TP + FP) (39)

4.1.2. Recall

Recall represents the model’s ability to correctly predict the positives out of actual positives. Thus, the formula to calculate the recall is given by equation (40),

Recall = True Positive/ (False Negative+True Positive) = TP / (FN + TP) (40)

4.1.3. F-*Measure*

F-Measure provides a way to combine both precision and recall into a single measure that captures both properties. The traditional F-measure is calculated as follows by equation (41),

F-Measure = 2* (Precision * Recall) / (Precision + Recall) (41)

4.1.4. Accuracy

Accuracy score represents the model’s ability to correctly predict both the positives and negatives out of all the predictions. Mathematically, it represents the ratio of sum of true positive and true negatives out of all the predictions by equation (42).

Accuracy = TP + TN/ (TP + TN + FP + FN) (42)

4.2. Results comparison

In this section shows the performance comparison results of the proposed FCNN classifier, existing Modified Convolutional Neural Network (MCNN) and Multi Layer Perceptron Neural Network (MLPNN) with three datasets has been shown in the table 4. The performance metrics values are given in the following Table 4.

Table 4. Performance Evaluation Results vs. Methods

Metrics	**Liver Disease - Methods**			**HCC - Methods**			**Gene expression data – methods**		
	MLPNN	**MCNN**	**FCNN**	**MLPNN**	**MCNN**	**FCNN**	**MLPNN**	**MCNN**	**FCNN**

Precision (%)	84.35	88.57	90.78	64.7059	74.3889	89.2308	59.5648	66.1326	71.2418
Recall (%)	59.85	94.11	97.36	57.4687	61.0499	85.4169	81.7352	82.2747	91.5966
F-Measure (%)	70.02	91.25	93.96	60.8729	67.8635	85.4289	68.9107	73.3258	80.1471
Accuracy (%)	86.70	90.75	92.48	75.7576	82.8283	85.8586	73.4219	88.0399	88.3268
Time (Seconds)	12.7738	2.9931	2.5479	8.9330	2.4384	2.0825	113.1338	77.4492	51.0988

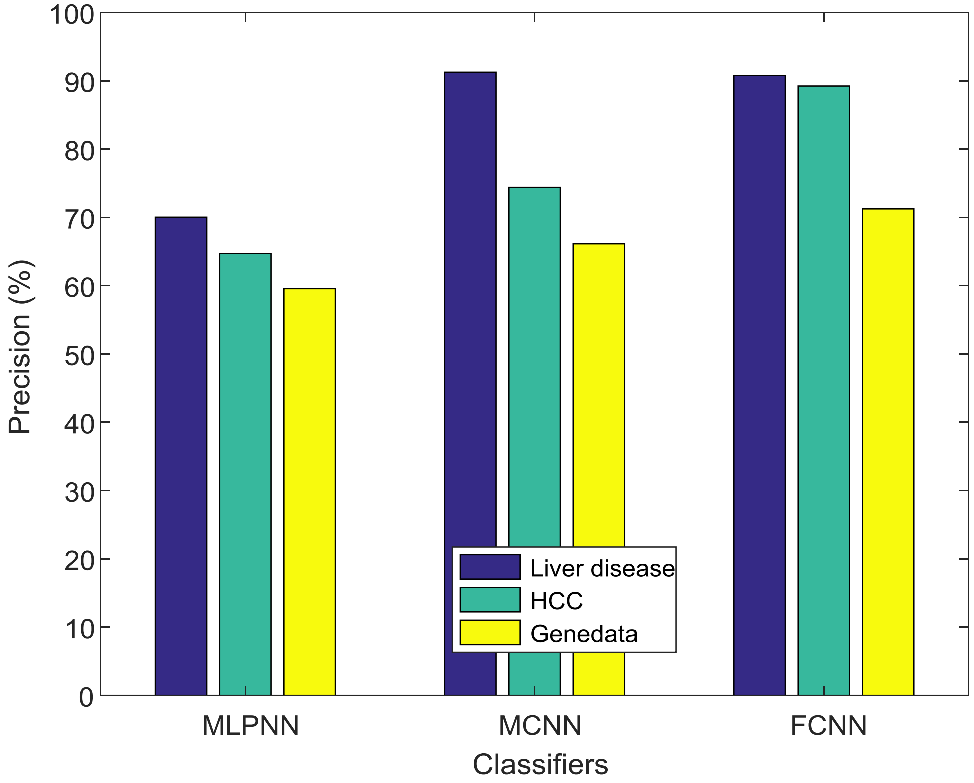

FIGURE 4. PRECISION RESULTS COMPARISON VS. METHODS

Figure 4 shows the performance comparison results of the precision metrics with respect to three different datasets such as liver disease and HCC of three classifiers such as MLPNN, MCNN and proposed FCNN classifier. From the results it concludes that the proposed FCNN classifier gives higher precision results of 90.78% for liver disease dataset, whereas other methods such as MLPNN and MCNN gives the precision results of 84.35% and 88.57% respectively (See Table 4). It is clearly observed from the graphical evaluation that the proposed system attains better Precision than the existing methods. This significant performance of proposed system is due to the modifications carried out in the feature selection and classification steps of the overall system.

Received 05 March 2021; Accepted 01 April 2021.

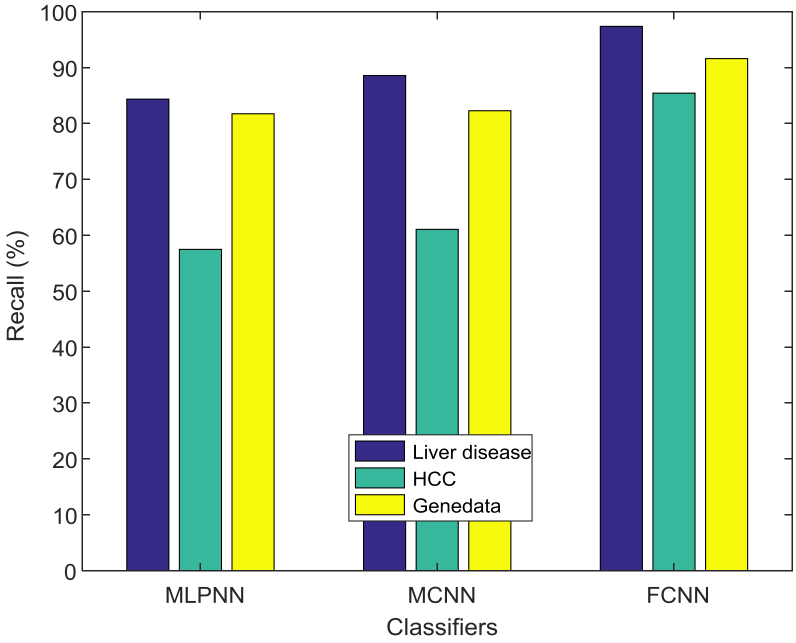

FIGURE 5. RECALL RESULTS COMPARISON VS. METHODS

Figure 5 shows the performance comparison results of the recall metrics with respect to three different datasets such as liver disease and HCC of three classifiers such as MLPNN, MCNN and proposed FCNN classifier. From the results it concludes that the proposed FCNN classifier gives higher recall results of 97.36% for liver disease dataset, whereas other methods such as MLPNN and MCNN gives the recall results of 59.85% and 94.11% respectively(See Table 4). This is mainly because of the inefficiency of the existing approach and it lacked the improvements carried out in the missing data imputation stage.

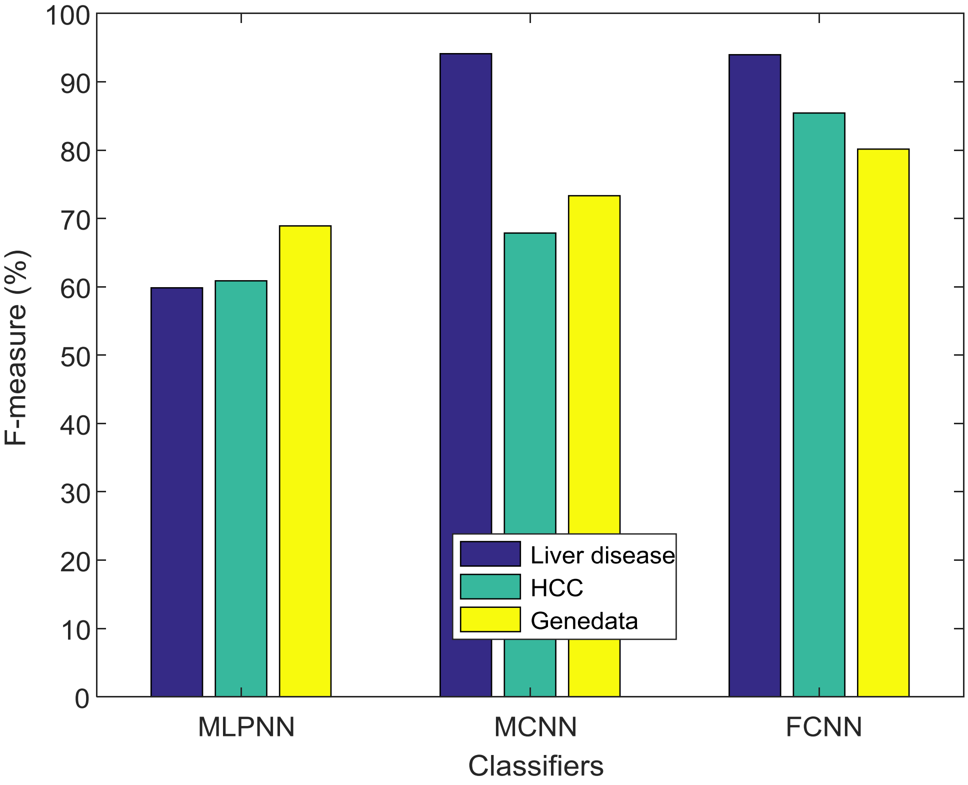

FIGURE 6. F-MEASURE RESULTS COMPARISON VS. METHODS

Received 05 March 2021; Accepted 01 April 2021.

Figure 6 shows the performance comparison results of the f-measure with respect to three different datasets such as liver disease and HCC of three classifiers such as MLPNN, MCNN and proposed FCNN classifier. From the results it concludes that the proposed FCNN classifier gives higher f-measure results of 93.96% for liver disease dataset, whereas other methods such as MLPNN and MCNN gives the f-measure results of 70.02% and 91.25% respectively(See Table 4). Thus, the performance of the proposed FCNN classifier is efficient and better when compared to the existing model.

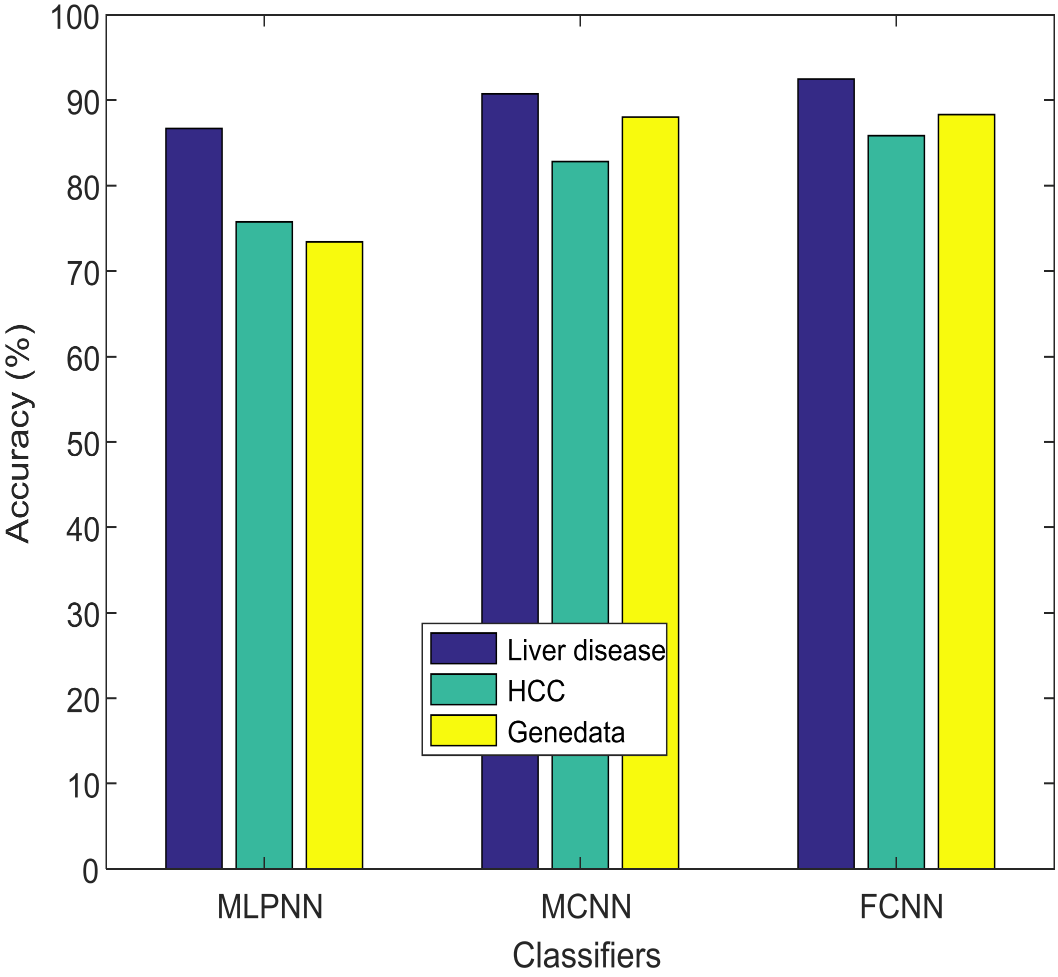

FIGURE 7. ACCURACY RESULTS COMPARISON VS. METHODS

Figure 7 shows the performance comparison results of the accuracy with respect to three different datasets such as liver disease and HCC of three classifiers such as MLPNN, MCNN and proposed FCNN classifier. From the results it concludes that the proposed FCNN classifier gives higher accuracy results of 92.48% for liver disease dataset, whereas other methods such as MLPNN and MCNN gives the accuracy results of 86.70% and 90.75% respectively(See Table 4). From this analysis it is proved that the proposed shows better performance than the existing technique. Proposed MCNN classifier shows improved increased accuracy than other methods. This is mainly because of the proposed formulations of fuzzy function in the CNN classifier.

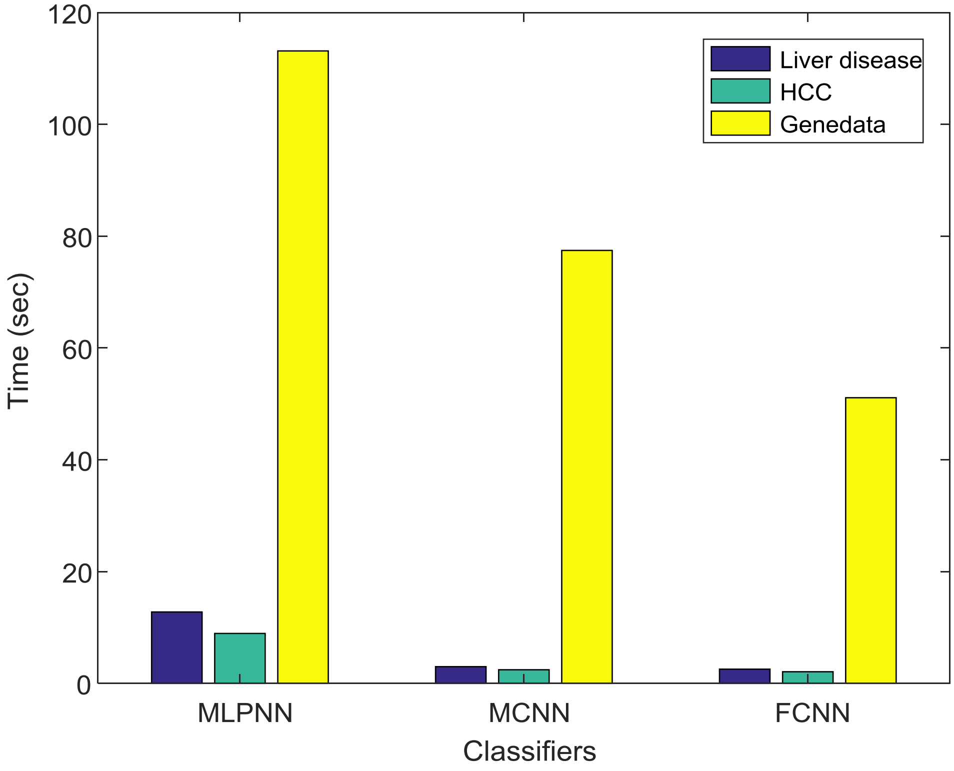

FIGURE 8. TIME RESULTS COMPARISON VS. METHODS

Figure 8 shows the time comparison results of three different datasets such as liver disease and HCC of three classifiers such as MLPNN, MCNN and proposed FCNN classifier. From the results it concludes that the proposed FCNN classifier has takes lesser time of 2.5479 seconds for liver disease dataset, whereas other methods such as MLPNN and MCNN has takes more time of 12.7738 seconds and 2.9931 seconds respectively(See Table 4).

5. CONCLUSION AND FUTURE WORK

Liver diseases are increasing phenomenally in the current world due to a variety of reasons. They are diagnosed very late where early predictions can save lives with proper treatment of the disease. MLTs have been used by many studies in the area of disease diagnostics in healthcare. These MLTs have issues in terms of accuracy and error rates in their proposed models. This work which uses data mining is a complete solution, working in three important phases namely pre-processing, feature selection and classifications. IFCM is used to pre-process liver data and replaces missing values by imputing them. The proposed MPCA is then used for reducing features (Dimensionality Reduction). The work uses Intelligent Ensembles operation for selecting optimal features where the SAFSA, MSBOA and SMFO techniques are used. MSBOA mimics butterflies food foraging behavior and uses mutation operator in its operations. Information Gain, Entropy and accuracy are considered as fitness values in this work's feature selections. The intelligent ensembles are combined in a voting scheme MVS for maximum optimality of feature selections. The combined selection of ensembles is then fed to FCNN for classifications which uses weights for optimizing its fuzzy membership function. 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. The performance results on UCI datasets show that FCNN is a better classifier than most other techniques use in classifications. Moreover, this works has achieved 90.78%, 97.36%, 93.96% and 92.48% in precision, recall, F-measure and accuracy values. In the future work, other classifiers will be implemented to increase the prediction rate of the system.

REFERENCES

1. Karthik, S., Priyadarishini, A., Anuradha, J. and Tripathy, B.K., 2011. Classification and rule extraction using rough set for diagnosis of liver disease and its types. AdvApplSci Res, 2(3), pp.334-345.

2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2018, 68, 394–424.

3. Fujiwara, N.; Friedman, S.L.; Goossens, N.; Hoshida, Y. Risk factors and prevention of hepatocellular carcinoma in the era of precision medicine. J. Hepatol. 2018, 68, 526–549.

4. Roth, G.A.; Abate, D.; Abate, K.H.; Abay, S.M.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; et al. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018, 392, 1736–1788.

5. Tseng, C.-H.; Hsu, Y.-C.; Chen, T.-H.; Ji, F.; Chen, I.-S.; Tsai, Y.-N.; Hai, H.; Thuy, L.T.T.; Hosaka, T.; Sezaki, H.; et al. Hepatocellular carcinoma incidence with tenofovir versus entecavir in chronic hepatitis B: A systematic review and meta-analysis. Lancet Gastroenterol. Hepatol. 2020, 5, 1039–1052.

6. Carrat, F.; Fontaine, H.; Dorival, C.; Simony, M.; Diallo, A.; Hezode, C.; De Ledinghen, V.; Larrey, D.; Haour, G.; Bronowicki, J.-P.; et al. Clinical outcomes in patients with chronic hepatitis C after direct-acting antiviral treatment: A prospective cohort study. Lancet 2019, 393, 1453–1464.

7. Stickel F, Hampe J. Genetic determinants of alcoholic liver disease. Gut 2012; 61: 150-159

8. Wang, S. & Summers, R. M. Machine learning and radiology. Medical image analysis 16, 933–951.

9. Ji, F., Liang, Y., Fu, S.-J., Guo, Z.-Y., Shu, M., Shen, S.-L., … Hua, Y.-P. (2016). A novel and accurate predictor of survival for patients with hepatocellular carcinoma after surgical resection: The neutrophil to lymphocyte ratio (NLR) combined with the aspartate aminotransferase/platelet count ratio index (APRI). BMC Cancer, 16(1), 137. https://doi.org/10.1186/s12885-016-2189-1

10. Wang, M., Wang, L., Ye, Z., & Yang, J. (2019). Ant lion optimizer for texture classification: A moving convolutional mask. IEEE Access, 7, 61697–61705

11. Storlie, C.B., Therneau, T.M., Carter, R.E., Chia, N., Bergquist, J.R., Huddleston, J.M. and Romero-Brufau, S., 2019. Prediction and inference with missing data in patient alert systems. Journal of the American Statistical Association, pp. 32-46.

12. Enders, C.K., Mistler, S.A. and Keller, B.T., 2016. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. Psychological methods, 21(2), p.222.

13. Romaniuk, H., Patton, G.C. and Carlin, J.B., 2014. Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods. American journal of epidemiology, 180(9), pp.920-932.

14. Aristiawati, K., Siswantining, T., Sarwinda, D. and Soemartojo, S.M., 2019, December. Missing values imputation based on fuzzy C-Means algorithm for classification of chronic obstructive pulmonary disease (COPD). In AIP Conference Proceedings (Vol. 2192, No. 1, p. 060003). AIP Publishing LLC.

15. Beni, G., 2020. Swarm intelligence. Complex Social and Behavioral Systems: Game Theory and Agent-Based Models, pp.791-818.

16. Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V. and Alonso-Betanzos, A., 2017. Ensemble feature selection: homogeneous and heterogeneous approaches. Knowledge-Based Systems, 118, pp.124-139.

17. Ma, X., Yang, Y., Tu, H., Gao, J., Tan, Y.T., Zheng, J.L., Bray, F. and Xiang, Y.B., 2016. Risk prediction models for hepatocellular carcinoma in different populations. Chinese Journal of Cancer Research, 28(2), p.150.

18. Kurosaki, M., Hiramatsu, N., Sakamoto, M., Suzuki, Y., Iwasaki, M., Tamori, A., Matsuura, K., Kakinuma, S., Sugauchi, F., Sakamoto, N. and Nakagawa, M., 2012. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. Journal of hepatology, 56(3), pp.602-608.

19. Cao, Y., Fan, J., Cao, H., Chen, Y., Li, J., Li, J. and Zhang, S., 2020. Prediction model for recurrence of hepatocellular carcinoma after resection by using neighbor2vec based algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, p.e1390.

20. Demir, F.B., Tuncer, T., Kocamaz, A.F. and Ertam, F., 2020. A survival classification method for hepatocellular carcinoma patients with chaotic Darcy optimization method based feature selection. Medical hypotheses, 139, p.109626.

21. Książek, W., Abdar, M., Acharya, U.R. and Pławiak, P., 2019. A novel machine learning approach for early detection of hepatocellular carcinoma patients. Cognitive Systems Research, 54, pp.116-127.

22. Dong, R.Z., Yang, X., Zhang, X.Y., Gao, P.T., Ke, A.W., Sun, H.C., Zhou, J., Fan, J., Cai, J.B. and 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), pp.3369-3374.

23. Zhang, Z.M., Tan, J.X., Wang, F., Dao, F.Y., Zhang, Z.Y. and Lin, H., 2020. Early diagnosis of hepatocellular carcinoma using machine learning method. Frontiers in bioengineering and biotechnology, 8, p.254.

24. Omran, D.A.E.H., Awad, A.H., Mabrouk, M.A.E.R., Soliman, A.F. and Aziz, A.O.A., 2015. Application of data mining techniques to explore predictors of HCC in Egyptian patients with HCV-related chronic liver disease. Asian Pacific Journal of Cancer Prevention, 16(1), pp.381-385.

25. Tian, X., Chong, Y., Huang, Y., Guo, P., Li, M., Zhang, W., Du, Z., Li, X. and Hao, Y., 2019. Using machine learning algorithms to predict hepatitis B surface antigen seroclearance. Computational and mathematical methods in medicine, Vol.2019, no. 6915850, pp.1-7.

26. https://archive.ics.uci.edu/ml/datasets/HCC+Survival

27. https://github.com/amazzocchi13/HCC-Prediction-Model-ML

28. Tang J.J., Zhang G.H., Wang Y.H., Wang H., Liu F. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. Transp. Res. Part C Emerg. Technol. 2015;51:29–40.

29. Huang, J., Mao, B., Bai, Y., Zhang, T. and Miao, C., 2020. An Integrated Fuzzy C-Means Method for Missing Data Imputation Using Taxi GPS Data. Sensors, 20(7), pp.1-19.

30. Eslam Ali Hassan, Ahmed Ibrahem Hafez, Aboul Ella Hassanien, and Aly A. Fahmy,2015, “Community Detection Algorithm Based on Artificial Fish Swarm Optimization”, Intelligent Systems'2014. Advances in Intelligent Systems and Computing, vol 323. Springer, Cham. https://doi.org/10.1007/978-3-319-11310-4_44

31. Arora, S.; Singh, S. Butterfly algorithm with Lèvy Flights for global optimization. In Proceedings of the 2015 International Conference on Signal Processing, Computing and Control (ISPCC), Waknaghat, India, 24–26 September 2015; pp. 220–224.

32. Arora, S. and Singh, S., 2019. Butterfly optimization algorithm: a novel approach for global optimization. Soft Computing, 23(3), pp.715-734.

33. Mirjalili, S., 2015. Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-based systems*, *89*, pp.228-249.

34. Fatih, A.B.U.T., AKAY, M.F. and 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), pp.3648-3664.

35. Williams, T. and Li, R., 2018, Wavelet pooling for convolutional neural networks. In International Conference on Learning Representations, pp.1-12.