Hepatitis C Stage Classification with hybridization of GA and Chi2 Feature Selection

Rukayya Umar[1], Steve Adeshina[2], Moussa Mahamat Boukar[3]
umarrukayya114@gmail.com, steve.adeshina@nileuniversity.edu.ng, musa.muhammed@nileuniversity.edu.ng
Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria[1,2,3]

Abstract— In metaheuristic algorithms such as Genetic Algorithm (GA), initial population has a significant impact as it affects the time such algorithm takes to obtain an optimal solution to the given problem. In addition, it may influence the quality of the solution obtained. In the machine learning field, feature selection is an important process to attaining a good performance model; Genetic algorithm has been utilized for this purpose by scientists. However, the characteristics of Genetic algorithm, namely random initial population generation from a vector of feature elements, may influence solution and execution time. In this paper, the use of a statistical algorithm has been introduced (Chi2) for feature relevant checks where p-values of conditional independence were considered. Features with low p-values were discarded and subject relevant subset of features to Genetic Algorithm. This is to gain a level of certainty of the fitness of features randomly selected. An ensembled-based learning model for Hepatitis has been developed for Hepatitis C stage classification. 1385 samples were used using Egyptian-dataset obtained from UCI repository. The comparative evaluation confirms decreased in execution time and an increase in model performance accuracy from 56% to 63%.

Keywords—*Hepatitis; Hepatitis prediction; Feature selection;*

I. INTRODUCTION

Feature selection is an important process for model development in the field of machine learning. A feature vector with n features has 2^n possible subsets of features; the goal of feature selection is to find which subset is best for the model performance. Several methods have been utilized for the task of feature selection, including Genetic algorithm, a metaheuristic algorithm which works on the principle of survival of the fittest. Generally, genetic algorithm has important steps, namely, creation of an initial population of individuals (randomly), evaluation of individuals fitness, selection of parent, performance of crossover of parent, mutation of population, repeatedly until individual is best. Each step in the process affects other, at the initial stage, individuals are generated randomly; researchers have demonstrated that the process poses difficulty obtaining a good initial population which eventually leads to increased computational time. In the domain of Feature selection, random generation of feature elements may equally lead to an increase in computational time. In this paper, we introduced the use of Chi-square as a step before the initial population generation, a statistical test applied to features to evaluate the likelihood of correlation or association between them using their frequency distribution. We then discard those irrelevant features and subject selected features to Genetic algorithm; doing this will give a level of certainty at the initial population generation and confidence in the fitness of each feature elements being randomly selected, thus minimizing execution cost and certainty of optimal solution at the end.

Genetic Algorithm, a commonly used optimization technique, has been applied in several different domains to solve complex problems. It is used to create robust methods by hybridization with other techniques. Hybridization of GA and Particle Swarm Optimization for Optimal reactive Power flow problem aiming at minimization of power loses, achieve results performed well in minimization outperforming the two methods individually, but slower than GA[1]. Authors in [2] reviewed applications of GA in economics specifically commodity forecast, where different variations and hybrids of GA have been used and performed good. Meanwhile, the work of Aibinu at el. in [3] focuses on issues associated with significant difference found in fitness values of chromosomes when the roulette wheel selection approach is used; they proposed a new technique for the reproduction process which clustering-based Genetic Algorithm with polygamy and dynamic population control. The application of their technique in robot route optimization problems produced better results as compared to existing techniques. Generally, the implementation of GA begins with a population of (random) genes of chromosomes. In k-means clustering, for a proper estimation of the number of clusters, GA is used; however, the genes of chromosomes randomly selected results in poor clustering. Authors in [4] proposed an Adaptive GA to improve on the initial population technique. Similarly, authors in [5] developed a new strategy for the initial population to solve the combinatorial optimization problem of TSP. In the domain of combinatorial optimization such as Feature Selection, authors in [6] proposed a new approach to initial population generation based on linear regression analysis of the problem. In this paper, considering Feature selection, generally, GA implementation starts with the creation of an

Manuscript received December 5, 2021
Manuscript revised December 20, 2021
https://doi.org/10.22937/IJCSNS.2022.22.1.23

initial population of chromosomes (randomly). The created population may contain poor fitness[6]. Hence leading to a long period of convergence to an optimal solution. To address that, we propose hybridization of GA with Chi2 for this purpose

With bioinformatics data, the combination of class imbalance, high dimensionality, small dataset size, and noisy data makes the classification task much more challenging. In addition, when building classification models, one faces the challenge that there is no single classifier that performs well in all scenarios. Thus, numerous classification approaches, such as ensemble learning methods, have been developed to address this problem successfully in a majority of situations. Ensemble learning frequently performs better than single base classifiers in performing classification tasks. The performance of an ensemble learner is influenced by two key factors: accuracy and diversity of base classifiers [7]. Throughout the data-mining and machine-learning literature, numerous ensemble methods have been proposed and among the most commonly-used ensemble techniques are: Bagging [8], Boosting [9] and Random Forest [10].

II. REVIEW OF LITERATURE

Generally, hepatitis diagnosis is done by a routine blood testing or during blood donation. There are a number of factors to diagnosing hepatitis virus, which makes the physician's job difficult. A physician usually makes judgments by assessing a patient's current test results and referring to the previous judgments made on other patients with similar conditions. The first Method largely depends on the physician's experience and ability to compare the current patient result with earlier patients. Conclusions are drawn based on previous judgements made on patients with similar symptoms by the physician[27]. Thus, a number of factors must be considered before passing judgements. Hence a tool is required to aid influence the physician's decision. Several works have been done with regard to Hepatitis diagnosis and classification. The authors in [38] presented a paper comparing Principal Component Analysis, Chi-square and Genetic Algorithm for Hepatitis disease Classification.

a. Hepatitis Disease

Inflammation of the liver is referred to as hepatitis, often caused by viral hepatitis. Scientists have identified 5 unique hepatitis viruses, identified by the letters A, B, C, D, and E. While all cause liver disease; they vary significantly. These 5 types are of greatest concern because of the burden of illness and death they cause and the potential for outbreaks and epidemic spread. In particular, Hepatitis B and Hepatitis C lead to chronic disease in hundreds of millions of people and, together, are the most common cause of liver cirrhosis and cancer. Hepatitis C Virus(HCV) is a major cause of liver-related disease [11]; it is known that hepatitis C is a life-threatening disease in the world, with about 150 million people affected [12]. About 70% of infected people develop a chronic infection that leads to liver diseases like cirrhosis, cancer, liver failure, and hepatocellular carcinoma (HCC)[13], [14]. Statistics have shown that the number of people living with HCV virus is increasing despite the existence of cure [12], [15]. Annually, 1.75 million people newly acquire Hepatitis C virus infection[16]. According to World Health Organization (WHO), in 2020, Nigeria alone has estimated that about 20 million people are chronically infected with hepatitis B and C. The number of cases is rising, and WHO has called for "hepatitis hepatitis-free future" targeting 2030. Globally, About 170 million have chronic HCV infection [17][18].

Chronic Hepatitis C (CHC) is a life-threatening health condition leading to fibrosis, cirrhosis and liver cancer. Insightful information of chronic hepatitis C stage in infected patients is essential for managing the disease. The examination of tissue specimens was very crucial to CHC evaluation process for stage identification[19]. However, this method has been reportedly expensive and invasive, time-consuming, and at risk of complications[19], [20]. Hence, researchers have immensely contributed to this regard by proposing cost-effective and time-efficient approaches to obtaining information on CHC stage.

In an effort to address issues associated with Hepatitis disease, researchers have employed the use of state-of-the–art methods to analyze, forecast and generate insight from the Hepatitis dataset to aid and influence decision making and policy formulation for health practitioners. Observatory Systems[21]–[23], Statistical modeling[18] and Machine learning approaches[24] have been quite successful in doing tasks of information visualization, classification, diagnosis, prediction of Infectious diseases.

Chun-Tao Wai and Joel K. Greenson [19] constructed a simple noninvasive index model to predict fibrosis and Cirrhosis using laboratory results and obtained validation tests of 0.88 and 0.94 of APRI AUC, respectively. In recent work UCI dataset. Meanwhile, Artificial neural networks was used by authors in [26] for a similar purpose. Further, a single-stage classification model and a multistage stepwise classification model based on Neural Networks, Decision Tree, Logistic Regression, and Nearest Neighborhood clustering, have been developed to predict an individual's liver fibrosis degree[27]. Again, Nahla and Sana [28] considered cohorts of 166 Egyptian children with CHC and developed a staging and fibrosis prediction system. They concluded that ML is an 'addition to non-invasive liver fibrosis prediction and staging methods in pediatrics'. [29] developed a machine learning-based algorithm to identify patients with chronic HCV infection using health insurance claims alone and compared with previously developed ICD-9 code-based algorithm. Onursal fezullah et al. described

how they applied multilayer neural network on hepatitis disease for diagnosis using approximation of sigmoid function and Levenberg Marquardt (ml) learning algorithm were used to obtain classification accuracy[30]. Generally, datasets characteristics pose challenges to most leaning algorithms to perform well. These characteristics may be noise, poor quality, high dimensionality etc. Feature Selection is a dimensionality reduction technique that is used to improve model performance accuracy as well as interpretability.

b. Feature Selection

FS is considered as either a single objective or multi-objective combinatorial optimization problem[31][32]. Datasets collected come with a number of dimensions(features) or attributes. These features are necessary for classification or prediction task as each feature (independent variable) have some degree of contribution to the target variable (dependent variable) however, when the dimension is much (course of dimensionality) becomes an issue for the model performance. Hence, the goal of the feature selection process is as follows: given a dataset n described by m features (m dimensions), find the minimum number of m which describes the dataset as much as the original set of attributes do[33].

Using relevant feature elements improves the predictive accuracy of classification algorithms, less learning time, and better interpretability. Approaches to feature selection include methods like Principal component analysis, which obtain new features by compressing feature dimensions to a smaller number of features containing only the principal components. Other approaches extract a subset of features from the feature vector.

C. Ensemble Learning

Every model is associated with some limitations, which sometimes results in and as a result it makes error on training sample[34]. Decision Decision-making is an important component of our daily activities. One primary task of machine learning is the construction of good models from datasets[35]. Datasets consists of feature vectors and are characterized by different limitations such as model complexity, noise, data imbalanced, high dimensionality etc. For a machine learning model to be accepted as an optimal hypothesis to a specific domain, it has to capture all these characteristics when performing classification tasks on the datasets. However, it is difficult for a machine learning model to capture all these characteristics; as a result, it makes error on training samples.

In the last years, ensemble learning methods have gained a significant attention from the scientific community in the area of machine learning and data mining. Machine learning ensemble method integrates multiple learning algorithms to achieve better predictive performance than could be achieved from individual learning algorithms alone[7]. It has experimentally been proven to provide substantially better performance than their single base learner[36]. They have been applied to different real-world problems. The idea behind the techniques is to combine a set of various prediction models to obtain a composite global model that produces reliable and accurate predictions. Subsequently, different ensemble learning algorithms and techniques have been proposed and applied in various classification and regression problems. Among which are face recognition, medical diagnosis such as disease classification and prediction jobs, financial forecasting and so on[37]. In ensemble learning, a set of learning algorithms called base learners are put together to perform a task using different approaches of either bagging, boosting or stacking techniques. Given a task, an ensemble of L individual learners (h1……. hL). Given a set of observations $X = \{x_i \in M\}$ and set of labels $Y = \{yi \in N\}$ and training examples $D = \{xi,yi\}$ as input, learn model M on D and get H as classifier[7].

III. RESEARCH METHODOLOGY

In this study, the methodology consists of two important phases. Figure 1 above shows how the flow of the model development. **Firstly**, dataset was acquired from UCI repository, perform data preprocessing, data cleaning; furthermore, the dataset was fed to Chi2 algorithm for feature significance/independence test, features with good fitness are selected and subject it to GA for optimal subset of feature selection; and attributes selection were carried out. **Secondly**, model development, stacking method of ensemble learning-based model has been proposed using three classical ML algorithms as base learners and a combiner or (meta-model) for a robust predictive model.

a. Genetic Algorithm

Genetic algorithm is a method often utilized for complex optimization problems based on genetic concept and has been applied in solving different problems in different domains. It has been used in machine learning for different purposes like Feature Selection (combinatorial optimization). An important aspect of machine learning

deals with selecting most relevant feature elements for model training in classification or prediction tasks.

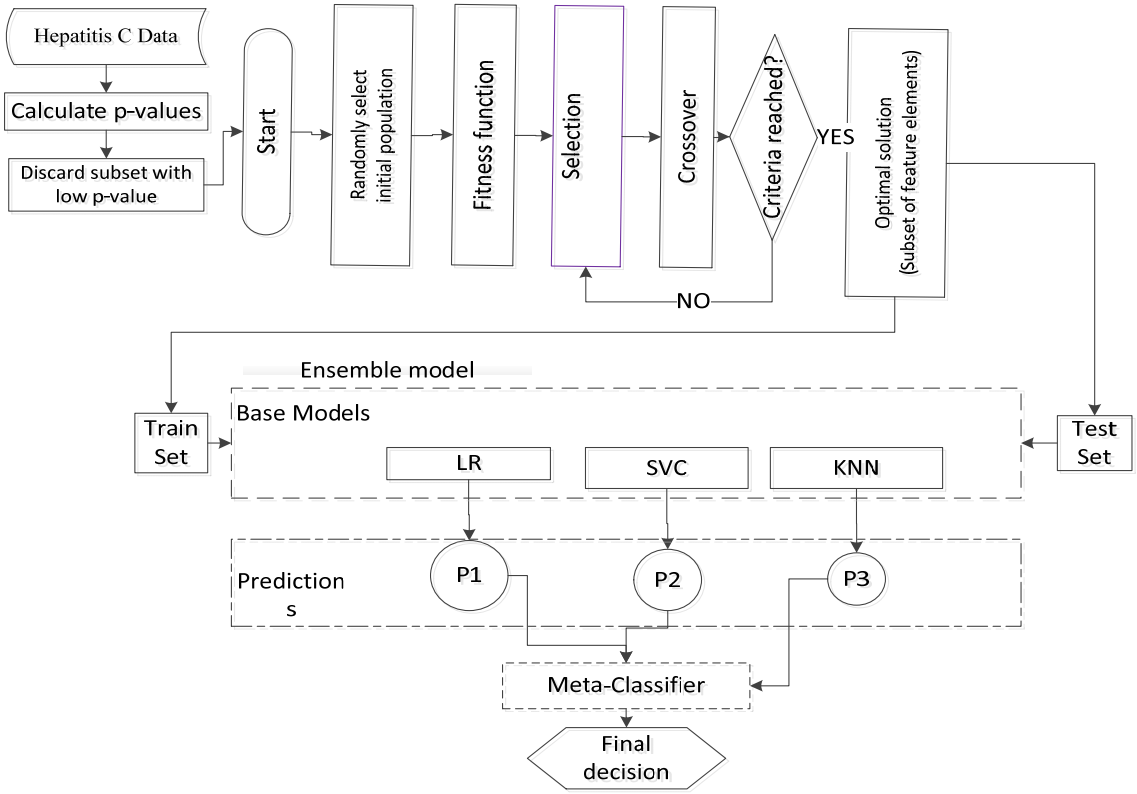


Figure 2. shows that the stage distribution in the datasets is evenly distributed, thus class balanced. Stage one (1) represents portal fibrosis, stage two (2) represents septa fibrosis, stage three (3) represents many septa, and stage four (4) represents Cirrhosis.

Genetic algorithm is an evolutionary algorithm commonly used in different domains. proposed in the history It consists of five main operations; Initial population generation, calculating the fitness of individual using Fitness function, Crossover, Mutation and Selection. In the selection phase, the fittest individuals that will pass genes to the next generation are selected. Strong selection leads to highly fit individuals forming the population. The commonly used methods used for selection are rank selection, tournament selection, roulette wheel, Boltzmann selection. However, the selection method is genetic algorithm comes with some limitations in for chromosome selection process. Certainty of feature elements selected is not guaranteed due to limitations associated with selection methods.

b. Chi-Square Algorithm

X^2 conducts significance test on the relationship between the values of a variable and category[41].

The significance is obtained using equation 1 below. The higher a chi-squared test score is the most likely to be independent and hence should be part of new set of features.

$$X^2 = \frac{A11 - E11}{E11} + \frac{A12 - E12}{E12} + \ldots \frac{Aij - Eij}{Eij}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(Aij - Eij)2}{Eij} \quad (1)$$

c. The Datasets

The dataset was obtained from UCI data Repository Machine Learning Databases. The dataset consists of 1385 samples with 29 attributes. The dataset was divided into four categories (classes) depending on the stage of the infection. There are 1385 observations with 29 features and target variables with to 4 stages (1, 2, 3, and 4). Table 1 shows the description of the data attributes, and figure 2 shows Hepatitis C (Class/Stage distribution).

Table 1. Data Variables

	Features		Features
1	Body Mass Index (BMI)	16	RNA 12
2	White blood cell (WBC)	17	RNA end-of-treatment
3	Red Blood Cells (RBC)	18	RNA Elongation Factor
4	Hemoglobin (HGB)	19	Baseline histological Grading
5	Platelets (Plat)	20	Age
6	Aspartate Transaminase Ratio (AST1)	21	Gender
7	Alanine Transaminase Ratio 1 week (ALT1)	22	Fever
8	Alanine Transaminase Ratio 12 weeks (ALT4)	23	Nausea/Vomiting
9	Alanine Transaminase Ratio 4 weeks (ALT12)	24	Diarrhea
10	Alanine Transaminase Ratio 24 weeks (ALT24)	25	Fatigue & generalized bone ache
11	Alanine transaminase Ratio 36 weeks (ALT36)	26	Jaundice
12	Alanine Transaminase Ratio 48 weeks(ALT48)	27	Headache
13	ALT after 24 w alanine transaminase ratio 24 weeks	28	Epigastric pain
14	RNA Base	29	Baseline histological staging (Class labels)
15	RNA 4		

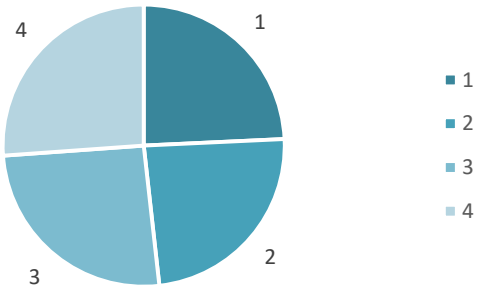


Figure 2. Stage (class) distribution chart.

IV. Evaluation Metrics

$$\text{Accuracy} = \frac{\sum_{i=1}^{l} TP_i + TN_i}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} (TP_i + FP_i)} \quad (2)$$

$$\text{Recall} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} TP_i + FN_i}$$

V. Experiments

In this study, two phases of experiment were carried out. With the aim to build a model that has high accuracy, precision and sensitivity with minimal feature elements. Two kay problems to tackle namely, classification of two Hepatitis C datasets acquired from UCI and feature selection optimization by hybridization of Chi2 with GA. In the first phase of the experiment, classification model was developed based on ensemble learning framework specifically using stacking approach. In this phase three classical machine learning algorithms namely, SVM, KNN and LR are the individual learners and LR the meta-classifier. The dataset was split into training and testing set, and GA was run on the dataset for Feature Selection. results were obtained from this experiment. In the second phase, the proposed approach has been implemented which is Hybridization of GA and Chi2 for FS. Chi2 was used to check the fitness of individual features. The created dataset is the used for classification task.

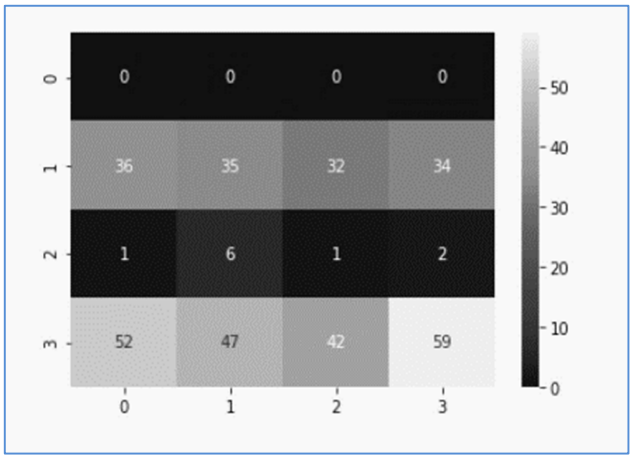
Algorithm, stacking ensemble

1. Input: training data D = $\{xi, yi\}_{i=1}^{k}$
2. Output: ensemble classifier M

3. Step 1: Learn base-level classifiers
4. For t = i to T do
5. Learn m_t based on D
6. end for
7. Step 2: construct new data set of predictions
8. For i = 1 to k do
9. Dm = {x'$_i$, yi}, where x'$_1$= {$m_1(x_i)$,.$mT(x_i)$}
10. end for
11. Step 3: learn meta-classifier
12. Learn M based on Dm
13. Return M

VI. RESULT ANALYSIS

The result obtained are in terms of the Evaluation metrics defined above including accuracy, precision and recall scores of ensemble model with all 29 features. The accuracy obtained is 52%, precision score 53%, recall 27%. The confusion matric is presented below:



Results of Feature Selection with Hybrid GA and Chi2

At this phase, Feature subset selection was performed usin g GA only, the following feature subsets were selected and the results obtained for test accuracy is: 0.59, Validation Ac curacy: 0.56, Individual: [1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0] with 15 feature subset s, namely,

Table 2: Optimal Features generated by GA FS

1	Age	9	Plat
2	Headache	10	AST 1
3	Diarrhea	11	ALT 48
4	Fatigue & generalized bone ache	12	RNA Base
5	Epigastric pain	13	RNA 12
6	WBC	14	RNNA 4
7	RBC	15	RNNA EOT
8	HGB	16	

The figure 3 below shows the p-values of feature elements, it is seen that some features are more relevant having high p -values and other with p-values. It is based on this we disca rd those feature elements with low p-values.

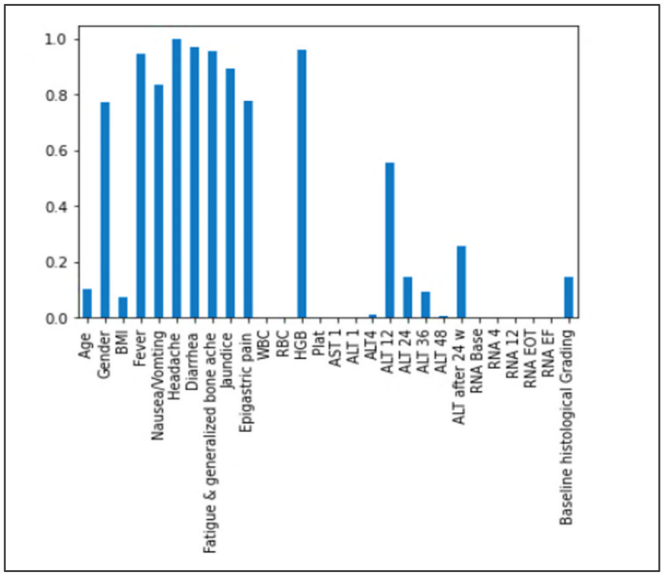


Figure 3. features shown based on their p-values

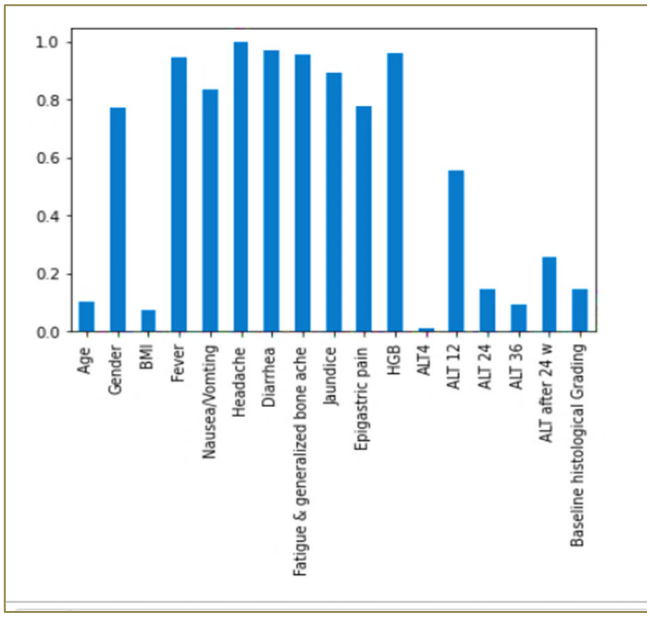


Figure 4. Feature elemesnts with good fitness.

In this figure, we have discarded the feature subset with very low p-values and the remaining feature subsets are shown.

The Hybrid GA and Chi2 is used for subset FS and the following was achieved Test accuracy: 66%, Validation Accuracy: 63% in a shorter period of time. Individual: [0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1] with 12 feature subsets presented in table 3.

Table 3: Optimal Features generated by GAChi2 FS

1	Gender	7	ALT 12
2	'BMI'	8	ALT 36
3	'Headache	9	ALT after 24 w
4	Fatigue & generalized bone ache	10	'Baseline histologicalGrading'
5	'Jaundice	11	HGB
6	'Epigastric pain	12	ALT4

The results obtained with HCV-Egyptian-data has no significant difference with all model developed, this work further use Hepatitis Disease dataset set to evaluate the effectiveness of the proposed FS approach.

VII. CONCLUSION

In this paper, the contribution findings are as follows: initially, a stacking ensemble model was implemented for Hepatitis C stage prediction and evaluated using Accuracy, Precision, recall metrics; using a hybrid of GA and Chi2 algorithm, we performed optimal Attributes Selection on HCV datasets. Under the same execution environment, the GA and GAChi2 have been implemented, ; our aim is to obtained a model with better performance accuracy than FS with GA alone with less execution period. The aim has been achieved, 56% accuracy has been obtained with GA Attributes selection and time taken, whereas GAChi2 achieved 63% accuracy within. Hence the result analysis of the data shows little significant difference. In the Future, GAChi2 will be applied to different datasets to test the method capability in different domains.

REFERENCES

[1] I. Cherki, A. Chaker, Z. Djidar, and N. Khalfallah, "A Sequential Hybridization of Genetic Algorithm and Particle Swarm Optimization for the Optimal Reactive Power Flow," 2019.

[2] K. Drachal, "A Review of the Applications of Genetic Algorithms to Forecasting Prices of Commodities," 2021.

[3] A. M. Aibinu, B. S. H, and M. N. C. M. Akachukwu, "A Novel Clustering based Genetic Algorithm (CGA) for Robot Route and Functions Optimization."

[4] X. Zhou, F. Miao, and H. Ma, "Genetic Algorithm with an Improved Initial Population Technique for Automatic Clustering of Low-Dimensional Data," pp. 1–23, 2018, doi: 10.3390/info9040101.

[5] Y. Deng, Y. Liu, and D. Zhou, "An Improved Genetic Algorithm with Initial Population Strategy for Symmetric TSP," vol. 2015, 2015.

[6] A. B. Hassanat, V. B. S. Prasath, M. A. Abbadi, S. A. Abu-qdari, and H. Faris, "An Improved Genetic Algorithm with a New Initialization Mechanism Based on Regression Techniques," doi: 10.3390/info9070167.

[7] T. G. Dietterich, "Ensemble methods in machine learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1857 LNCS, pp. 1–15, 2000, doi: 10.1007/3-540-45014-9_1.

[8] R. Richman and M. V. Wüthrich, "Nagging predictors," *Risks*, vol. 8, no. 3, pp. 1–26, 2020, doi: 10.3390/risks8030083.

[9] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," *Proc. 13th Int. Conf. Mach. Learn.*, pp. 148–156, 1996, doi: 10.1.1.133.1040.

[10] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780429469275-8.

[11] V. M. Cowton, J. B. Singer, R. J. Gifford, and A. H. Patel, "Predicting the effectiveness of hepatitis C virus

neutralizing antibodies by bioinformatic analysis of conserved epitope residues using public sequence data," *Front. Immunol.*, vol. 9, no. JUN, pp. 1–14, 2018, doi: 10.3389/fimmu.2018.01470.

[12] WHO, "WHO Global Hepatitis Report," 2017, [Online]. Available: http://apps.who.int/iris/bitstream/10665/255016/1/9789241565455-eng.pdf?ua=1.

[13] CDC, "Hepatitis C," *Osp. Magg.*, p. 2, 2015, doi: 10.1016/j.disamonth.2014.04.002.

[14] FIND, "Strategy for Hepatitis C 2015-2020," 2014.

[15] C. W. Shepard, L. Finelli, and M. J. Alter, "Global epidemiology of hepatitis C virus infection.," *Lancet. Infect. Dis.*, vol. 5, no. 9, pp. 558–67, 2005, doi: 10.1016/S1473-3099(05)70216-4.

[16] *Global hepatitis report, 2017*. 2017.

[17] A. Roos *et al.*, "Investigations, findings, and follow-up in patients with chest pain and elevated high-sensitivity cardiac troponin T levels but no myocardial infarction," *Int. J. Cardiol.*, vol. 232, no. June, pp. 111–116, 2017, doi: 10.1016/j.ijcard.2017.01.044.

[18] M. Jefferies, B. Rauff, H. Rashid, T. Lam, and S. Rafiq, "Update on global epidemiology of viral hepatitis and preventive strategies," *World J. Clin. Cases*, vol. 6, no. 13, pp. 589–599, 2018, doi: 10.12998/wjcc.v6.i13.589.

[19] C. T. Wai *et al.*, "A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C," *Hepatology*, vol. 38, no. 2, pp. 518–526, 2003, doi: 10.1053/jhep.2003.50346.

[20] P. Halfon *et al.*, "Accuracy of hyaluronic acid level for predicting liver fibrosis stages in patients with hepatitis C virus," *Comp. Hepatol.*, vol. 4, pp. 1–7, 2005, doi: 10.1186/1476-5926-4-6.

[21] R. Tinati, X. Wang, I. Brown, T. Tiropanis, and W. Hall, "A Streaming Real-Time Web Observatory Architecture for Monitoring the Health of Social Machines," *Proc. 24th Int. Conf. World Wide Web - WWW '15 Companion*, pp. 1149–1154, 2015, doi: 10.1145/2740908.2743977.

[22] R. Umar, A. David, and A. Adesiyun, "Observatory system for monitoring hepatitis c development in Nigeria," *2019 15th Int. Conf. Electron. Comput. Comput. ICECCO 2019*, no. Icecco, pp. 1–6, 2019, doi: 10.1109/ICECCO48375.2019.9043245.

[23] A. H. Observatory, "Health Observatories," no. April, 2016.

[24] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 570–579, 2012, doi: 10.1016/j.cmpb.2011.08.003.

[25] T. M. Ghazal *et al.*, "Hep-pred: Hepatitis C staging prediction using fine gaussian SVM," *Comput. Mater. Contin.*, vol. 69, no. 1, pp. 191–203, 2021, doi: 10.32604/cmc.2021.015436.

[26] D. Sarma *et al.*, "Artificial Neural Network Model for Hepatitis C Stage Detection," *EDU J. Comput. Electr. Eng.*, vol. 1, no. 1, pp. 11–16, 2020, doi: 10.46603/ejcee.v1i1.6.

[27] A. M. Hashem, M. E. M. Rasmy, K. M. Wahba, and O. G. Shaker, "Single stage and multistage classification models for the prediction of liver fibrosis degree in patients with chronic hepatitis C infection," *Comput. Methods Programs Biomed.*, vol. 105, no. 3, pp. 194–209, 2012, doi: 10.1016/j.cmpb.2011.10.005.

[28] N. H. Barakat, S. H. Barakat, and N. Ahmed, "Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach," *Healthc. Inform. Res.*, vol. 25, no. 3, pp. 173–181, 2019, doi: 10.4258/hir.2019.25.3.173.

[29] M. A. Khan, J. E. Soh, M. Maenner, W. W. Thompson, and N. P. Nelson, "A machine-learning algorithm to identify hepatitis C in health insurance claims data," *Online J. Public Health Inform.*, vol. 11, no. 1, pp. 98–99, 2019, doi: 10.5210/ojphi.v11i1.9685.

[30] D. M. Journal, "An application of multilayer neural network on hepatitis disease diagnosis using approximations of sigmoid activation function Hepatitis disease dataset," vol. 42, no. 2, pp. 150–157, 2015, doi: 10.5798/diclemedj.0921.2015.02.0550.

[31] W. Mostert and K. M. Malan, "Comparative Analysis," pp. 1–16, 2021.

[32] S. Wu, Y. Hu, W. Wang, X. Feng, and W. Shu, "Application of Global Optimization Methods for Feature Selection and Machine Learning," vol. 2013, 2013.

[33] P. L. Lanzi, "Fast feature selection with genetic algorithms: A filter approach," *Proc. IEEE Conf. Evol. Comput. ICEC*, pp. 537–540, 1997, doi: 10.1109/icec.1997.592369.

[34] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.

[35] "Ensemble Methods, Foundations and Algorithms.pdf." .

[36] A. K. Seewald, "Towards a theoretical framework for ensemble classification," *IJCAI Int. Jt. Conf. Artif. Intell.*, no. 3, pp. 1443–1444, 2003.

[37] P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," *Algorithms*, vol. 13, no. 6, 2020, doi: 10.3390/A13060140.

[38] R. Umar, M. M. Boukar, S. Adeshina, and S. Dane, "Machine Learning Approaches for Optimal Parameter Selection for Hepatitis Disease Classification."

[39] F. E. H. Tay and L. Shen, "A modified Chi2 algorithm for discretization," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 666–670, 2002, doi: 10.1109/TKDE.2002.1000349.

[40] H. Liu and R. Setiono, "Feature Selection via Discretization," vol. 9, no. 4, pp. 1995–1998, 1997.

[41] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," *Proc. Int. Conf. Tools with Artif. Intell.*, pp. 388–391, 1995, doi: 10.1109/tai.1995.479783.